Math 2: Ordinary Differential Equations

Tom Hutchcroft

PMA, Caltech.
Email: t.hutchcroft@caltech.edu

Last updated at 23:41 on Saturday 7th December, 2024

Warning: These notes are an *incomplete draft*! Please proceed with caution, and do not hesitate to let me know if you find any mistakes or if you find any of the explanations unclear. The easiest way to do this is to send me an email.

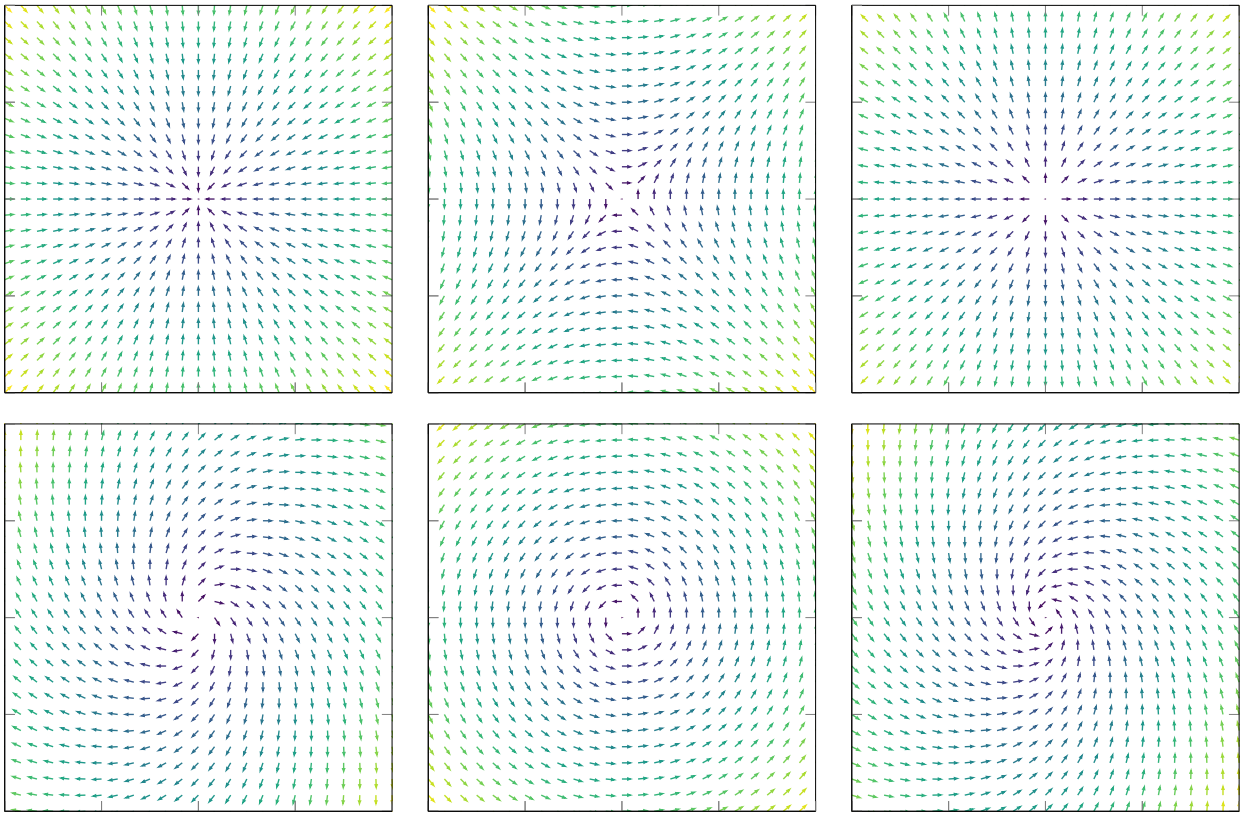


Figure 1: Six possible local behaviours near a fixed point for a first-order autonomous equation in two dimensions. Clockwise from top left: Sink, Saddle, Source, Spiral Sink, Center, Spiral Source.

Preamble

You are reading the lecture notes for the course *Math 2 - Analytical* as taught at Caltech in Fall 2024. The "analytic" track is the more theory-focused of the two introductory courses on differential equations taught at Caltech. It is intended primarily for sophomore students who have already taken introductory proof-based courses on (multivariable) calculus and linear algebra.

This is an unusual course! Most universities will offer an ODE course along the lines of Caltech's Math 2 (practical), and might also offer advanced undergraduate / graduate courses on dynamical systems (in which one typically studies the qualitative properties of ODE solutions without actually *solving* the ODEs, often because doing so is prohibitively difficult); basic ODE theorems like the existence and uniqueness of solutions would only be covered tangentially in real analysis courses. This may be one of the only courses in the world that 1. is for undergraduates, 2. focusses primarily on *quantitative* aspects of ODEs, and 3. is fully rigorous. Because the course is so unusual, there are not really any good texts for it, which is why we will rely so heavily on these notes. **Please let me know if you find mistakes, or anything is unclear: do not wait until the surveys at the end of the quarter to complain!**

I think this is a great course covering many extremely interesting and important topics that are often omitted from an undergraduate education in mathematics – I hope by the end you will agree with me!

Organisation of these notes. These notes will sometimes contain more detail than will be given in the lectures (and certainly much more text!).

Timeline of lectures.

Lecture 1: Introduction. Review of limits, continuity, differentiability, etc. Discussion of some simple examples. Pages 5-9.

Lecture 2: Further discussion of examples, including examples with non-uniqueness of solutions to IVPs. Brief discussion of IVPs vs. BVPs. Finite-time blow-up. Pages 9-15.

Lecture 3: Definition of linear ODEs. Linear algebra recap. Pages 18-21.

Lecture 4: Norms on vector spaces. Continuity, differentiability etc. for functions between vector spaces. Open sets in vector spaces. Pages 21-26.

Lecture 5: Vector-valued ODEs. Phase space and reduction to first-order. Started discussing matrix representation of linear ODEs. Pages 27-33.

Lecture 6: Calculus with matrices. Defined and proved some basic properties of matrix exponentials. Pages 33-37.

Lecture 7: Finished discussing basic properties of matrix exponentials and relations to ODEs. Started talking about how to do computations with matrix exponentials. Pages 37-42.

Lecture 8: Computing matrix exponentials for non-diagonalizable matrices via Jordan normal form. The damped spring equation. The cookbook solution to constant coefficient linear ODEs. Pages 43-50.

Lecture 9: The big-picture strategy behind the proof of Picard-Lindelof. Completeness of the space of continuous functions with the uniform norm. Pages 52-55.

Lecture 10: Continuity of integration. Term-by-term differentiation. Proof of (local, first-order) Picard-Lindelöf. Pages 56-60.

Lecture 11: nth order local Picard-Lindelof. Definition of locally space-Lipschitz functions. Statement of global Picard-Lindelof and proof of the glueing lemma. Pages 60-63.

Lecture 12: Proof of global Picard-Lindelof. Consequences for autonomous ODEs. Automatic smoothness of solutions to smooth ODEs. Solution to first-order linear ODEs. Pages 64-66.

Lecture 13: Gronwall's lemma. Started discussing continuous dependence of solutions on initial conditions. Pages 66-69.

Lecture 14: Finished proof of continuous dependence of solutions on initial conditions. Continuous dependence of solutions on coefficients. Pages 70-71.

Lecture 15: Inhomogeneous linear ODEs and Duhamel's principle. Started discussing separable ODEs. Pages 71-77.

Lecture 16: Finished discussing separable ODEs, started discussing Laplace transforms. Pages 77-82.

Lecture 17: Injectivity of the Laplace transform. More transformation rules. Laplace transforms of products of polynomials, exponentials, and trig functions. Pages 82-85.

Lecture 18: More Laplace transforms; solving constant coefficient linear ODEs using Laplace transforms and partial fractions. Computing the Laplace transform of the solution to a non-constant coefficient linear ODE. Started discussing convolutions. Pages 86-91.

Lecture 19: Finished discussion of convolutions. Started section on formal power series and formal operations on power series. Pages 92-97.

Lecture 20: Formal solutions to ODEs. Formal and function operations coincide within the radius of convergence. Computation of some examples. Pages 98-102.

Lecture 21: Existence of formal solutions; convergence of formal solutions to polynomial ODEs in standard form. Some examples where things go wrong for more general polynomial ODEs. Started discussing ordinary generating functions. Pages 103-108 and 111-112. (Forgot to discuss composition of formal power series and analytic ODEs...)

Lecture 22: (Formal) ordinary generating functions and their applications to combinatorial problems. Started discussing exponential generating functions. 112-117.

Lecture 23: Continued discussing (formal) exponential generating functions and their applications to combinatorial problems. Difference equations. Composition of formal power series and formal solutions to analytic ODEs. Pages 118-122 and 110-111.

Lecture 24: Asymptotic notation. Asymptotic expansions. Computed the asymptotic expansion of the exponential integral.

Lecture 25: Overview of Tauberian theory and functions of regular variation.

Lecture 26: Series solutions beyond power series; the Frobenius method.

Lecture 27: A worked example of the Frobenius method; some heuristic methods to guess the first term in the series expansion for more general ODEs.

Lecture 28: A brief introduction to the local analysis of autonomous ODEs near equillibrea.

Contents

1 Introduction

This course is an introduction to the study of *ordinary differential equations* (ODEs), i.e., equations among the derivatives of a function taking a single variable as input, often thought of as time, but which may have a multiple-variable output. The word 'ordinary' distinguishes such equations from *partial differential equations* (PDEs), which concern the (partial) derivatives of multivariable functions and are typically much more difficult to study. For functions $f : \mathbb{R} \to \mathbb{R}^d$, the most general form[1] such an ODE can take is

$$F\left(t; f(t), \frac{df}{dt}, \frac{d^2f}{dt^2}, \ldots, \frac{d^nf}{dt^n}\right) = 0 \tag{1.1}$$

for some given function F taking $1 + (n+1)d$ variables (equivalently, one one-dimensional variable and $n+1$ d-dimensional variables) for some $n \geq 1$. Such an equation is called an **order** n ODE; the problem is to solve for f (either on $\mathbb{R}$ or some appropriate interval), given some appropriate boundary conditions. Note that part of what is means for f to solve the ODE (1.1) on some open interval is for f to be n-times differentiable on that interval[2]. In practice we will almost always consider the not-very-restrictive special case where our ODE is of the form

$$\frac{d^nf}{dt^n} = F\left(t; f(t), \frac{df}{dt}, \frac{d^2f}{dt^2}, \ldots, \frac{d^{n-1}f}{dt^{n-1}}\right). \tag{1.2}$$

This is partly because ODEs arising in applications almost always have this form (or can be written in this form), and partly because it is for equations of this form that the basic existence and uniqueness theorems are formulated (as we will see, uniqueness can fail for simple algebraic reasons in the more general setting of (1.1)).

It should go without saying that ODEs are ubiquitous throughout engineering and the natural sciences. In these notes I will assume that you already have a reason to care about ODEs from some other aspect of your life[3] (or are happy to proceed unmotivated) and focus on the mathematical aspects. Even within mathematics, ODEs arise in a huge variety of contexts and, especially once one moves beyond the linear setting, require many distinct tools to solve.

Since different kinds of ODE all have their own distinct characteristics and personality, it

[1] Of course one is not obligated to call the independent variable t.

[2] In fact one can formulate precise notions of what it means for a not-necessarily-differentiable function to satisfy a differential equation using what are called *weak derivatives*. This perspective is particularly important in PDE. Unfortunately, the theory of weak derivatives requires some measure theory to formulate and is outside the scope of this course. If you are interested, the Wikipedia page https://en.wikipedia.org/wiki/Weak_solution is a good place to start.

[3] In the 2023/2024 versions of the notes I have been adding discussions of examples arising in applications. These are meant to add flavour only; it's not a problem if you're not familiar with the examples discussed. For even more flavour you can try making your favourite AI chatbot turn each of your homework problems into a word problem.

is easy for an introductory course to take on a disjointed, ad hoc, or "cookbook" style in an effort to quickly teach the student how to solve a wide variety of examples. Students inclined towards a more theoretical perspective are likely to find such a presentation uninspiring, and perhaps to (wrongly) deduce that the subject is not interesting. This could not be further from the truth! Indeed, the late 19th / early 20th century real analysis, functional analysis, measure theory etc. that you have learned and will learn in your other math classes – with all its attendant εs and δs – was developed with the primary goal of putting the theory of differential equations on a rigorous footing, and the theory of differential equations (particularly PDEs) remains one of the most research-active areas of theoretical mathematics to this day. Moreover, although ODEs are usually thought of as a "settled topic" in contrast to PDEs, there are still many aspects of the theory that are not completely worked out. Indeed, the 16th of *Hilbert's problems*, a highly influential list of 23 problems posed by David Hilbert in 1900 as a challenge for 20th century mathematics, concerns ODEs and remains open to this day. (Besides this, there are also, of course, settled parts of the theory that are much too advanced to be covered in this course!) Partly this is all a matter of branding, with many aspects of the topic of contemporary research interest now coming under the heading of *"dynamical systems"* rather than ODE per se.

We hope that the presentation of the basic theory of ODEs given in these notes will help those same theory-inclined students appreciate the mathematical beauty of the topic, and later on, perhaps, to better appreciate the work of their colleagues in analysis and applied mathematics (if they are not working in these topics themselves). Of course there must be a trade-off to developing the underlying theory at greater length than usual, which will most likely be that we will have much less time to go through worked examples in class and will cover fewer computational solution techniques. Since I do still want you to develop your facility solving concrete equations, the problem sets will likely have a more computational character than the lectures themselves.

1.1 Differentiability Recap I

Before proceeding further, let us quickly recall some important definitions from Math 1. We write $\mathbb{R}$ for the set of all real numbers. Given two real numbers $a \leq b$, we write (a, b) for the open interval $(a, b) = \{x \in \mathbb{R} : a < x < b\}$ and $[a, b]$ for the closed interval $[a, b] = \{x \in \mathbb{R} : a \leq x \leq b\}$. Similarly one can define the half-open and intervals $(a, b]$ and $[a, b)$, noting that (a, b), $(a, b]$, and $[a, b)$ are all empty if $a = b$. One can also consider open intervals with endpoints at $-\infty$ or $+\infty$, so that e.g. $(-\infty, b) = \{x \in \mathbb{R} : x < b\}$ and $(-\infty, +\infty) = \mathbb{R}$. A set of real numbers $I \subseteq \mathbb{R}$ is said to be an **interval** if $[a, b] \subseteq I$ for every $a, b \in I$: Every non-empty interval is of the form (a, b), $[a, b]$, or $(a, b]$ for some $-\infty \leq a < b \leq +\infty$. We say that an interval I is **non-trivial** if it contains at least two distinct points.

A sequence of real numbers $(x_n)_{n\geq 1}$ is said to **converge** to a real number x if for every $\varepsilon > 0$ there exists $N < \infty$ such that $|x - x_n| \leq \varepsilon$ for every $n \geq N$. We write "$x_n \to x$ as $n \to \infty$" as a shorthand to mean that $(x_n)_{n\geq 1}$ converges to x. Given an interval I and a

function $f : I \to \mathbb{R}$, we say that f is **continuous** at a point $x \in I$ if $f(x_n) \to f(x)$ as $n \to \infty$ whenever $(x_n)_{n\geq 1}$ is a sequence in I converging to x as $n \to \infty$. We say that f is continuous on the interval I if it is continuous at every point of I. We say that f is **differentiable** at a point x of a non-trivial interval I if there exists a real number $f'(x)$ such that

$$\frac{f(x_n) - f(x)}{x_n - x} \to f'(x)$$

whenever $(x_n)_{n\geq 1}$ is a sequence in $I \setminus \{x\}$ converging to x. (Note that if x is a left or right endpoint of I then this is equivalent to what is usually called left- or right-differentiability as appropriate.) Intuitively this means that f is approximated to first order by a straight line of slope $f'(x)$ when we zoom in near x. We say that f is differentiable on I if it is differentiable at every point of I and that f is **continuously differentiable** on I if its derivative f' defines a continuous function on I. Note that differentiable functions are always continuous.

Exercise 1. Give an example of a function $f : [-1, 1] \to \mathbb{R}$ that is differentiable on $[-1, 1]$ but for which the derivative f' is not continuous at 0.

Differentiation is **linear**, meaning that if $f, g : I \to \mathbb{R}$ are differentiable at $x \in I$ and $a, b \in \mathbb{R}$ then $af + bg$ is differentiable at x with derivative $af' + bg'$. The **product rule** states that if $f, g : I \to \mathbb{R}$ are differentiable at $x \in I$ then their product fg is also differentiable at x with derivative $f'g + fg'$. The **chain rule** states that if $f : I_1 \to I_2$ and $g : I_2 \to \mathbb{R}$ are such that f is differentiable at $x \in I_1$ and g is differentiable at $f(x) \in I_2$ then the composition $g \circ f : I_1 \to \mathbb{R}$ (defined by $g \circ f(x) = g(f(x))$) is differentiable at x with derivative $f'(x)g'(f(x))$.

We say that a function $f : I \to \mathbb{R}$ is **twice differentiable** if it is differentiable and its derivative f' is also differentiable. The derivative of f' is called the **second derivative** of f and is written as f'', $f^{(2)}$, or using Leibniz's notation as e.g. $\frac{d^2 f}{dx^2}$ or $\frac{d^2 f}{dt^2}$ (depending on what one wants to call the input variable). We can similarly define n-times differentiability and the nth derivative $f^{(n)}$ for each $n \geq 1$, and say that a function $f : I \to \mathbb{R}$ is **smooth** (a.k.a. infinitely differentiable) on I if it is n-times differentiable for every $n \geq 1$. If f is n-times differentiable on an interval I, Taylor's approximation theorem states that

$$f(x) = f(x_0) + (x - x_0)f'(x_0) + (x - x_0)^2 \frac{f''(x_0)}{2!} + \cdots + (x - x_0)^n \frac{f^{(n)}(x_0)}{n!} \pm o\left(|x - x_0|^n\right)$$

as $x \to x_0$ in I, where $\pm o(|x - x_0|^n)$ represents a function $h(x)$ satisfying $h(x)/(x - x_0)^n \to 0$ as $x \to x_0$ in I and $n!$ (a.k.a. n factorial) denotes the product of the first n positive integers. More formally, this means that if $x_0 \in I$ and $(x_m)_{m\geq 1}$ is a sequence in $I \setminus \{x_0\}$ converging to x_0 then

$$\frac{f(x_m) - \left(f(x_0) + (x_m - x_0)f'(x_0) + (x_m - x_0)^2 \frac{f''(x_0)}{2!} + \cdots (x_m - x_0)^n \frac{f^{(n)}(x_0)}{n!}\right)}{(x_m - x_0)^n} \to 0$$

as $m \to \infty$.

Exercise 2. Let $I \subseteq \mathbb{R}$ be a non-trivial interval and let $n \geq 1$. Prove that $f : I \to \mathbb{R}$ is n-times differentiable on I if and only if there exists an open interval $\tilde{I}$ containing I and an n-times differentiable function $\tilde{f} : \tilde{I} \to \mathbb{R}$ such that $\tilde{f}(t) = f(t)$ for every $t \in I$. (NB[4]: The claim is trivial if I is already open.)

1.2 Lessons from some simple ODEs

In this section we will solve a few of the most simple ODEs and discuss how the behaviour of the solutions will be reflected in more complicated examples. We begin with the simplest ODE of all.

Lemma 1.1. *Let $I \subseteq \mathbb{R}$ be a non-trivial interval and let $f : I \to \mathbb{R}$ be a differentiable function. Then f satisfies the first order ODE*

$$\frac{df}{dt} = 0$$

for every $t \in I$ if and only if there exists a constant $C \in \mathbb{R}$ such that $f(t) = C$ for every $t \in I$.

Proof. Constant functions obviously have zero derivative, so it suffices to prove conversely that every function with zero derivative is constant. Since f is differentiable on I, the mean-value theorem states that for every $a < s < t < b$ there exists $s \leq x \leq t$ such that $f(t) - f(s) = (t - s)f'(x)$. Since $f'(x) = 0$ for every $x \in I$ it follows that $f(t) = f(s)$ for every $s < t$ in I, implying the claim. □

Lemma 1.2. *Let $I \subseteq \mathbb{R}$ be a non-trivial interval, let $g : I \to \mathbb{R}$ be continuous and let $f : I \to \mathbb{R}$ be a differentiable function. For each $t_0 \in I$, f satisfies the first order ODE*

$$f'(t) = g(t) \tag{1.3}$$

for every $t \in I$ if and only if there exists a constant $C \in \mathbb{R}$ such that

$$f(t) = \int_{t_0}^{t} g(s)\,\mathrm{d}s + C$$

for every $t \in I$.

Proof. This is just the fundamental theorem of calculus! Let's nonetheless take a moment to unpack this a little since there are a few different possible statements of the fundamental theorem. The most basic statement is that if $g : I \to \mathbb{R}$ is continuous and $t_0 \in I$ then

$$\frac{d}{dt}\int_{t_0}^{t} g(s)\,\mathrm{d}s = g(t)$$

[4] "NB" is an abbreviation for the Latin term "Nota Bene," which translates to "Note Well." It's traditionally used in academic and formal writing to emphasize an important point or detail that the reader should not overlook.

for every $t \in I$. Since constant functions have zero derivative and differentiation is linear, this implies that every function of the form $\int_{t_0}^t g(s)\,\mathrm{d}s + C$ has derivative $g(t)$ for every $t \in I$. On the other hand, if $f(t)$ is any function satisfying $\frac{d}{dt}f(t) = g(t)$ for every $t \in I$ then $\frac{d}{dt}(f(t) - \int_{t_0}^t g(s)) = 0$ for every $t \in I$, so that there exists a constant $C \in \mathbb{R}$ such that $f(t) - \int_{t_0}^t g(s) = C$ for every $t \in I$ by Lemma 1.1. This is equivalent to the claim. □

Lemma 1.3. *Let $I \subseteq \mathbb{R}$ be a non-trivial interval, let $g : I \to \mathbb{R}$ be continuous and let $f : I \to \mathbb{R}$ be a differentiable function. Then f satisfies the first order ODE*

$$f'(t) = g(t)f(t) \tag{1.4}$$

for every $t \in I$ if and only if

$$f(t) = f(t_0)\exp\left[\int_{t_0}^t g(s)\,\mathrm{d}s\right] \tag{1.5}$$

for every $t, t_0 \in I$.

The easy way to see that the solutions to (1.4) should be of the form (1.5) is to think in terms of the logarithmic derivative $(\log f)' = f'/f$. Indeed, *if f is positive on I*, then the ODE (1.4) is equivalent to the ODE

$$(\log f)' = g(t),$$

which we can solve using Lemma 1.2 and recover the solution (1.5). Generally it is a good idea to "think logarithmically" whenever the rate of change of a quantity is described as a proportion of its current value as in (1.4), since in such cases the logarithm of the quantity will often satisfy a simpler ODE than the original quantity. This perspective is particularly natural in applications related to e.g. population growth, epidemics, compound interest, and so on. At the level of generality of Lemma 1.3 this approach raises some annoying issues regarding negative or zero values of f, which have to be treated by case analysis, so we will follow a slightly different approach.

Proof. We can easily verify from the chain rule and the fundamental theorem of calculus that functions of the form (1.5) satisfy (1.4):

$$\begin{aligned}\frac{d}{dt}\left(A\exp\left[\int_0^t g(s)\,\mathrm{d}s\right]\right) &= A\frac{d}{dt}\exp\left[\int_0^t g(s)\,\mathrm{d}s\right] \\ &= A\exp\left[\int_0^t g(s)\,\mathrm{d}s\right]\frac{d}{dt}\int_0^t g(s)\,\mathrm{d}s = g(t)A\exp\left[\int_0^t g(s)\,\mathrm{d}s\right].\end{aligned}$$

Conversely, suppose that $f : I \to \mathbb{R}$ is any solution to (1.4). Then we have by the product

rule that

$$\frac{d}{dt}\left(\exp\left[-\int_0^t g(s)\,\mathrm{d}s\right] f(t)\right) = \frac{d}{dt}\exp\left[-\int_0^t g(s)\,\mathrm{d}s\right] f(t) + \exp\left[-\int_0^t g(s)\,\mathrm{d}s\right]\frac{d}{dt}f$$
$$= -g(t)\exp\left[-\int_0^t g(s)\,\mathrm{d}s\right] f(t) + g(t)\exp\left[-\int_0^t g(s)\,\mathrm{d}s\right] f(t) = 0$$

for every $t \in I$ and hence by Lemma 1.1 that there exists a constant $A \in \mathbb{R}$ such that

$$\exp\left[-\int_0^t g(s)\,\mathrm{d}s\right] f(t) = A$$

for every $t \in I$. This is equivalent to the claim. □

Note that in each of these three simple examples, every solution of the relevant first-order ODE is completely determined by its value at a single point. This is in fact a very general phenomenon. We state the relevant principle imprecisely for now; a precise statement will be given later as the *Picard-Lindelöf Theorem*.

Principle 1. *If f satisfies a "nice" first-order ODE, then it's value at every point in time is determined by its value at a single point in time. In particular, a "nice" first-order ODE describing functions taking values in a d-dimensional space should have a "d-dimensional" space of solutions.*

Having such a uniqueness statement will be very useful: it means we can often guess a family of solutions, easily verify that they satisfy the ODE (since differentiation is much easier than integration), and be guaranteed from general principles that this family includes all other solutions.

We will return to the meaning of the word "nice" later in the course. To convince you that some restrictions are necessary, let us give a simple example of an ODE that is not "nice" in the sense that Principle 1 does not apply to it.

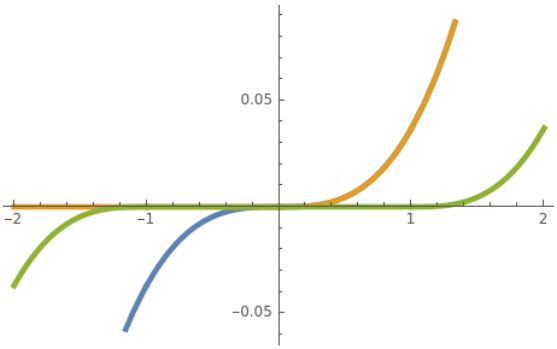
Example 1.4. Consider the ODE

$$f'(t) = f(t)^{2/3}.$$

We can verify that one solution of this ODE is given by $f(t) = \frac{1}{27}t^3$:

$$f'(t) = \frac{1}{9}t^2 = \left(\frac{1}{27}t^3\right)^{2/3} = f(t)^{2/3}$$

(Later we'll see how to solve this kind of equation without guessing: It's an example of a *separable* ODE.) In particular, this solution has $f(0) = 0$. *This is not the only solution with*



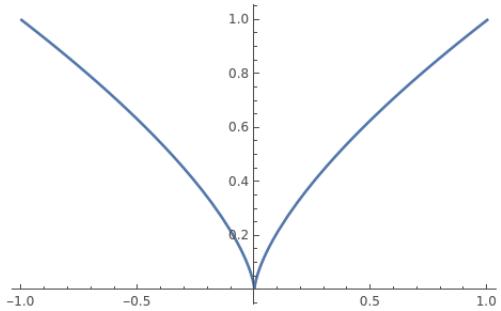


Figure 2: Left: Three different solutions to the same first-order ODE $f' = f^{2/3}$ all with $f(0) = 0$. The blue and orange curves are identical for $t \geq 0$. Right: We will see that this ODE is "not nice" partly because the function $f \mapsto f^{2/3}$ used in its definition has a cusp at 0, where its derivative converges to $\pm\infty$ as we approach from the left or the right.

this property! Indeed, the constant function $f(t) \equiv 0$ is also a solution. Moreover, for each $-\infty \leq t_- < t_+ \leq \infty$ the function defined piecewise by

$$f(t) = \begin{cases} \frac{1}{27}(t - t_+)^3 & t \geq t_+ \\ 0 & t_- \leq t \leq t_+ \\ \frac{1}{27}(t - t_-)^3 & t < t_- \end{cases}$$

is a solution to the same ODE, with $f(0) = 0$ if $t_- \leq 0 \leq t_+$ (in particular this function is differentiable). We will later prove that these are the only solutions to the ODE.

A real-world example with non-unique solutions

Suppose that a cylindrical tank has a small hole in its side. *Torricelli's law* states that if the tank is filled up to a height h above the hole at time zero, then the height $h(t)$ of the water after t seconds satisfies the ODE

$$h'(t) = -\frac{\text{area of hole}}{\text{cross-sectional area of cylinder}} \sqrt{\mathbb{1}(h \geq 0) 2gh(t)},$$

where g is the constant for acceleration under gravity. Similarly to above, this ODE has solutions

$$h(t) = \frac{1}{4}(t_0 - t)^2 \mathbb{1}(t \leq t_0)$$

for $t_0 \in \mathbb{R}$ and

$$h(t) = c$$

for each $c \leq 0$. In particular, there are infinitely many different solutions with $h(0) = 0$. This makes perfect sense physically: If we see the tank at some time and the water is filled only to the height of the hole, we should not be able to guess how long ago any additional water emptied from the tank.

What about higher-order ODEs? As in the first-order case, the simplest examples admit simple solutions using the fundamental theorem of calculus.

Lemma 1.5. *Let $I \subseteq \mathbb{R}$ be a non-trivial interval, let $n \geq 1$, let $g : I \to \mathbb{R}$ be continuous and let $f : I \to \mathbb{R}$ be an n-times differentiable function. For each $t_0 \in I$, f satisfies the nth order ODE*

$$\frac{d^n f}{dt^n} = g(t) \tag{1.6}$$

for every $t \in I$ if and only if there exist constants $C_0, \ldots, C_{n-1} \in \mathbb{R}$ such that

$$f(t) = \frac{C_{n-1}}{(n-1)!}t^{n-1} + \frac{C_{n-2}}{(n-2)!}t^{n-2} + \cdots + C_1 t + C_0 + \int_{t_0}^{t}\int_{t_0}^{s_0}\int_{t_0}^{s_1}\cdots\int_{t_0}^{s_{n-2}} g(s_{n-1})\,\mathrm{d}s_{n-1}\,\mathrm{d}s_{n-2}\cdots\mathrm{d}s_0 \tag{1.7}$$

for every $t \in I$.

Remark 1.6. Taking $g \equiv 0$, it follows that the only functions f satisfying $\frac{d^n f}{dt^n} = 0$ on a non-trivial interval are polynomials of degree at most $n-1$.

Proof. Fix $t_0 \in I$. Let $g_0 = g$ and for each $n \geq$ let g_n be the nth antiderivative of g defined recursively by

$$g_n(t) = \int_{t_0}^{t} g_{n-1}(s)\,\mathrm{d}s = \int_{t_0}^{t}\int_{t_0}^{s_0}\int_{t_0}^{s_1}\cdots\int_{t_0}^{s_{n-2}} g(s_{n-1})\,\mathrm{d}s_{n-1}\,\mathrm{d}s_{n-2}\cdots\mathrm{d}s_0$$

so that, by the fundamental theorem of calculus,

$$\frac{dg_n}{dt} = g_{n-1}$$

for every $n \geq 1$ and $t \in I$. It follows in particular that the nth derivative of g_n is equal to g, and since the nth derivative of a degree $n-1$ polynomial is zero it follows that every function of the form (1.7) is a solution to (1.6).

It remains to prove conversely that *every* solution to (1.6) is of the required form. We will prove this by induction on $n \geq 1$, the base case $n = 1$ having already been treated in Lemma 1.2. Let $n > 1$, suppose that the claim has already been proven for every smaller value of n, and suppose f is n-times differentiable and solves (1.6) for some $n > 1$. Rewriting the equation (1.6) as

$$\frac{d}{dt}\left(\frac{d^{n-1}f}{dt^{n-1}}\right) = g(t)$$

and applying Lemma 1.2 yields that

$$\frac{d^{n-1}f}{dt^n} = g_1(t) + C$$

for some constant $C \in \mathbb{R}$ and every $t \in I$. It follows by the induction hypothesis that f is of the form $f(t) = P(t) + G(t)$ where P is a polynomial of degree at most $n-1$ and $G(t)$ is the $(n-1)$th antiderivative of $g_1 + C$. Since integration is linear, $G(t)$ is equal to $g_n + \frac{C}{(n-1)!}t^{n-1}$ and f has the required form. □

The first main takeaway from this example is that solving an nth order ODE is, in general, at least as difficult as integrating a function n times. Often it is much more difficult! Expressing the solution to an ODE in terms of standard functions and their (iterated) integrals is known as **solution by quadrature**. Note that while integrals of 'standard' functions do not always have exact expressions in terms of other 'standard' functions, solving an ODE by quadrature is about as good as we can hope for in many situations, and much better than we can hope for in others. While we will see throughout the course that there a few important classes of equations that can always be solved by quadrature (separable equations, first order linear equations, higher order linear equations with constant coefficients...), you should not be fooled into thinking that ODEs can "usually" be solved by quadrature.

Note that in this example, knowing the value of a solution at a single point was not sufficient to determine the rest of the solution, but knowing the value of the first $n-1$ derivatives of the function at a single point did suffice to determine the solution. Again, this turns out to be a very general principle.

Principle 2. *If f satisfies a "nice" nth-order ODE, then it's value at every point in time is determined by the value of f and its first $n-1$ derivatives at a single point in time. In particular, a "nice" nth order ODE describing a real-valued function should have an "n-dimensional" space of solutions.*

We will see later that nth-order ODEs can always be thought of as first-order ODEs in a higher dimension, so that in fact Principle 2 is a consequence of Principle 1. It is worth pointing out again how *useful* a precise uniqueness theorem of this form would have been for solving the equation (1.6): We could simply have verified that the solutions of the desired form solve the ODE, and deduce that these are the only solutions by an easy application of a general uniqueness result.

IVPs and BVPs. When solving ODEs in practice we are usually interested in finding *all* solutions of the ODE only insofar as it helps us understand the *specific solution arising in our problem.* Since the space of all solutions of a "nice" nth order ODE will be n-dimensional (in some sense), we should expect to need to specify the values of n parameters in order to pin down any specific solution. (That is, there should be n degrees of freedom when we specify a solution.) The simplest way to do this is via what's called an **initial value problem** (IVP), where we specify the value of the function and its first $n-1$ derivatives at some point in time (often $t_0 = 0$): principle 2 says that when the ODE is sufficiently "nice", the IVP should always have a unique solution. (Moreover, we will see that the precise meaning of "nice" – the local Lipschitz property – is not a very restrictive condition.)

In applications one also often encounters **boundary value problems**, where some data of a solution is specified at *both* endpoints of an interval. For example, in a second-order ODE we might specify the values of the function at both endpoints of an interval and not specify the derivative anywhere. This still gives two degrees of freedom, so it *might* be the case that we have existence and uniqueness of solutions. Unlike for IVPs, however, it is no longer at all the case that any sufficiently nice BVP has a unique solution. One can easily come up with very simple, well-behaved examples where there are either no solutions or multiple solutions: In the second-order case above with the interval $I = [0, 1]$ (and assuming that the IVP has a unique solution for every initial value and derivative) this comes down precisely to whether, for each initial condition $f(0) = x_0$ and initial derivative $f'(0) = \lambda$, the map sending λ to the value of $f(1)$ in the solution of the relevant IVP is a bijection $\mathbb{R} \to \mathbb{R}$. There is simply no reason for this to be the case in many examples.

Exercise 3. In this exercise we will study IVPs and BVPs for the ODE $f'' = -f$. Let $I \subseteq \mathbb{R}$ be a non-trivial interval.

1. (Existence of solutions to the IVP.) Prove that for each $t_0 \in I$, $x_0 \in \mathbb{R}$, and $\lambda \in \mathbb{R}$, there exist constants $A, \sigma \in \mathbb{R}$ such that $f(t) = A\cos(t - \sigma)$ is a solution to the ODE $f'' = -f$ with $f(t_0) = x_0$ and $f'(t_0) = \lambda$.

2. (Uniqueness of solutions to the IVP.) Prove that the function $A\cos(t - \sigma)$ you found in step 1 is the *only* solution to the IVP with $f(t_0) = x_0$ and $f'(t_0) = \lambda$. (Hint: Show that the derivative of f/g is identically zero whenever f and g are arbitrary solutions of the IVP.)

3. (Existence and/or uniqueness of solutions to the BVP depends on the choice of I and the boundary data.) Show that if $I = [0, \pi/2]$ then for every $x_0, x_1 \in \mathbb{R}$ there exists a unique function $f : [0, \pi/2] \to \mathbb{R}$ solving the ODE $f'' = -f$ with $f(0) = x_0$ and $f(\pi/2) = x_1$. What happens when $I = [0, \pi]$ or $I = [0, 2\pi]$? (Hint: Use the multiple angle formula to write the solutions in the form $A\cos(t) + B\sin(t)$ instead of $A\cos(t - \sigma)$.)

A ball in a bowl

Imagine that a ball (modelled as a point) sits inside a bowl. If we write down an equation describing the shape of the sides of the bowl, the location of the ball will satisfy a second-order ODE, with the *acceleration* of the ball determined by the two forces acting on it: gravity and the contact force. (The basic point of this example still stands if we include things like friction; the state of the system is still described by a second-order ODE.) If we push the ball from the center of the bowl with a certain velocity, its trajectory is a solution to the initial value problem where we fix the location and velocity of the ball at time zero. On the other hand, we could try to study the *boundary value problem*, where we are given the ball's location (but not its velocity) at two times, and want to determine its trajectory in between. In this example we easily see that the solution to this BVP is typically not unique. In particular, if the ball is in the center of the bowl at two times, this could either be because we didn't push it at all, or pushed it with exactly the right velocity that it rolled back down to the center at the second time. There will in fact be more than two solutions in general, corresponding to the ball rolling back and forth multiple times before passing through the center of the bowl at some later time.

BVPs from variational problems

Boundary value problems often arise in the analysis of *variational problems*, where a curve is chosen to maximize some function (often the negative of the energy) with its initial and final values fixed. The fact that solutions to such problems are often described by ODEs is a consequence of the *Euler-Lagrange equations*. For example, when we hold the two ends of a string in the air and let the rest of the string come to rest, the string will settle in a way that minimizes its potential energy subject to the constraints we have placed on the locations of its endpoints. Using the Euler-Lagrange equations, one can deduce that the function describing the string satisfies a second-order ODE, and more specifically a BVP: We know where the string is at its two endpoints but not the derivative of the curve at either endpoint. Look up the term "catenary" to learn more about this particular example.

The importance of linearity. All the examples we have looked at so far are **linear**. This means that there exist functions $a_0, \dots, a_{n-1} : I \to \mathbb{R}$ and $b : I \to \mathbb{R}$ such that our differential equation can be expressed in the form

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

Linear ODEs are called **homogeneous** if $b(t) \equiv 0$ and **inhomogeneous** otherwise. Linear equations are typically much easier to solve than nonlinear equations, and their solutions often have underlying linear-algebraic content. For example, we will see that constant-coefficient

linear ODEs can always be solved by computing the Jordan normal form of an associated matrix. One basic principle about these equations, that we have already seen illustrated in our examples above, is that the solutions to a *homogeneous* linear ODE form a vector space under pointwise addition and scalar multiplication, while for inhomogeneous linear ODEs we have that

$$\left\{ f : \frac{d^n f}{dt^n} + a_{n-1}(t)\frac{d^{n-1} f}{dt^{n-1}} + \cdots + a_0(t) f(t) = b(t) \right\}$$
$$= \left\{ f + f_0 : \frac{d^n f}{dt^n} + a_{n-1}(t)\frac{d^{n-1} f}{dt^{n-1}} + \cdots + a_0(t) f(t) = 0 \right\}$$

for every solution f_0 to the original inhomogeneous ODE $\frac{d^n f}{dt^n} + a_{n-1}(t)\frac{d^{n-1} f}{dt^{n-1}} + \cdots + a_0(t) f(t) = b(t)$. This means that we can always find the general solution to an inhomogeneous linear ODE by finding any single solution (e.g. by guessing), often called the *particular solution*, and finding the general solution to the associated homogeneous ODE, which is easier.

A simple nonlinear example. Before moving on, let us consider a simple example of a *nonlinear* ODE, which will also exhibit the phenomenon of *finite-time blow-up*.

Lemma 1.7. *Let I be a non-trivial interval. A differentiable function $f : I \to \mathbb{R}$ satisfies the ODE*

$$f' = f^2$$

if and only if $f(t) = 0$ for all $t \in I$ or there exists a constant $t_0 \notin I$ such that

$$f(t) = \frac{1}{t_0 - t}$$

for every $t \in I$.

Proof. As usual, we can easily check that functions of the form $f(t) = 0$ and $f(t) = 1/(t_0 - t)$ satisfy the relevant ODE by calculus. For uniqueness, suppose that $f : I \to \mathbb{R}$ is not identically zero and satisfies the ODE $f' = f^2$ on I. For the uniqueness part of the proof, we would like to say that either $f \equiv 0$ or the reciprocal function $1/f$ satisfies

$$\left(\frac{1}{f}\right)' = -\frac{f'}{f^2} = -1,$$

so that we can conclude using Lemma 1.2. The problem is that this doesn't make sense when f is zero. To get around this problem we will use a trick that will come up many times in the course: We analyze the solution on a (possibly) smaller "good" interval in which the technique we want to use works, then use the output of this analysis to show that this interval is actually the whole domain of the solution.

Suppose that f is not identically zero, and let t_1 be such that $f(t_1) \neq 0$. Since I is

non-trivial and f is continuous, the quantities

$$t_- = \sup\{t \leq t_1 : t \notin I \text{ or } t \in I \text{ and } f(t) = 0\}$$
$$t_+ = \inf\{t \geq t_1 : t \notin I \text{ or } t \in I \text{ and } f(t) = 0\}$$

satisfy $t_- < t_+$. Let $\tilde{I} = (t_-, t_+) \subseteq I$, so that f is non-zero and satisfies $f' = f^2$ when restricted to $\tilde{I}$. (This is the "good" interval alluded to above.) On $\tilde{I}$, the reciprocal function $1/f$ satisfies

$$\left(\frac{1}{f}\right)' = -\frac{f'}{f^2} = -1$$

so that, by Lemma 1.2, there exists a constant C such that

$$\frac{1}{f} = C - t$$

for every $t \in \tilde{I}$. By definition of $\tilde{I}$, each endpoint of $\tilde{I}$ must therefore either be an endpoint of I or be equal to C, and the claim follows by continuity of f on I. □

This example exhibits **finite-time blowup**: There are solutions on intervals $I \subsetneq \mathbb{R}$ that cannot be extended to solutions defined on the whole real line, because the function converges to infinity as t converges to some t_0. Of course this can also occur because we put a singularity directly into the ODE, such as in the ODE

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

the solutions of which will always have a singularity at 1. In contrast, the location of the singularity in a solution to $f' = f^2$ depends on the choice of solution.

The ODE $f' = f^2$ is also an example of an **autonomous** equation, i.e., an ODE of the form

$$f^{(n)}(t) = F\left(f(t), f'(t), \ldots, f^{(n-1)}(t)\right)$$

where the right hand side has no direct dependence on t. We will study autonomous equations extensively later on in the course. Note that if $f(t)$ is a solution to an autonomous ODE then so is the shifted function $f(t - t_0)$ for every $t_0 \in \mathbb{R}$.

Remark 1.8. The solutions of the ODE $f' = f^2 + 1$ are all of the form $f(t) = \tan(t - t_0)$, and hence always have two singularities a distance π apart from each other, but with the location of these singularities depending on the choice of solution. The fact that these are the *only* solutions to the equation will follow from Picard-Lindelöf (which we will state and prove later in the notes).

Finite-time blow-up in the real world

You might guess that solutions with finite-time blow-up should not be relevant to real-world phenomena, but this is not true. For example, ODEs whose solutions have finite-time blowup often arise in the study of phase transitions, where the finite-time blow-up represents a critical point at which a rapid change or transition occurs. In the context of phase transitions, this can be understood as a moment where the system undergoes a rapid change from one state to another. Here, the finite-time blow-up is not a mathematical anomaly but rather a (good approximation to a) physical reality that needs to be addressed and understood. It often represents some quantity changing by a factor of order the number of particles in the system (e.g. changing from an electron-Volt-order quantity to a Joule-order quantity), which is taken to be infinite in the mathematical models where these ODEs arise.

To give a specific example, the ODE $f' = f^2$ that we just studied arises in the study of *branching processes*, which can be used as e.g. a simple model of the spread of a disease, where the parameter we're differentiating with respect to is "R_0", the average number of people that an infected person infects. In this set-up, $f(R_0) = 1/(1 - R_0)$ is the average total number of infected individuals when we start with a single infected individual. The finite-time blow-up represents the phase transition between the disease dying off quickly when $R_0 < 1$ to infecting a very large number of people (an infinite number of people in the mathematical model) when $R_0 > 1$.

Exercise 4. Let $I \subseteq \mathbb{R}$ be a non-trivial interval, let $g : I \to \mathbb{R}$ be continuous, let $f : I \to \mathbb{R}$ be a differentiable function, and let $a \in \mathbb{R}$ be a constant. Prove that f solves the first-order ODE $f' = af + g$ if and only if

$$f(t) = e^{a(t-t_0)} f(t_0) + e^{a(t-t_0)} \int_{t_0}^{t} e^{-a(s-t_0)} g(s) \, \mathrm{d}s$$

for every $t, t_0 \in I$. (Hint: Find an ODE satisfied by $e^{-at} f$.)

Exercise 5. Let I be a non-trivial interval and let $\alpha > 1$. Prove that a differentiable function $f : I \to (0, \infty)$ satisfies the ODE

$$f' = f^\alpha$$

if and only if there exists a constant $t_0 \geq \sup I$ such that

$$f(t) = \left(\frac{1}{(\alpha - 1)(t_0 - t)} \right)^{1/(\alpha-1)}$$

for every $t \in I$.

Exercise 6. Let $I \subseteq \mathbb{R}$ be a non-trivial interval, let $g : I \to \mathbb{R}$ be continuous, and let

$f : I \to \mathbb{R}$ be a differentiable function. Prove that f satisfies the ODE

$$f' = e^{g-f}$$

if and only if for each $t_0 \in I$ there exists $A > 0$ such that

$$f(t) = \log \left[\int_{t_0}^{t} e^{g(s)} \, \mathrm{d}s + A \right]$$

for every $t \in I$.

2 Constant coefficient linear ODEs

In this section we begin to study ODEs in a more systematic manner. We will, for now, be concerned primarily with **linear** ODEs, i.e., ODEs of the form

$$f^{(n)} + a_{n-1}(t)f^{(n-1)} + \cdots + a_1(t)f'(t) + a_0(t)f(t) = b(t).$$

for some functions $a_0, \ldots, a_{n-1}$ and $b : I \to \mathbb{R}$, where I is a non-trivial interval. We recall that a linear ODE is called **homogeneous** when $b \equiv 0$ and **inhomogeneous** otherwise. Since linear algebra plays an important role in the study of linear ODE, we begin by recapping some of the basic theory that we'll need.

2.1 Linear Algebra Recap

Linear algebra will play an extremely important role throughout our study of differential equations. While we expect you to have already taken a course on linear algebra, we quickly review some of the most important facts here before moving forward.

Vector spaces. A (real) **vector space** $(V, +, \cdot)$ is a set V equipped with addition and scalar multiplication operations

$$\begin{aligned} + : V \times V &\longrightarrow V & \qquad \cdot : \mathbb{R} \times V &\longrightarrow V \\ (x, y) &\longmapsto x + y & (a, x) &\longmapsto ax \end{aligned}$$

satisfying the following axioms:

1. (Associativity of vector addition) $x + (y + z) = (x + y) + z$ for all $x, y, z \in V$.
2. (Commutativity of vector addition) $x + y = y + z$ for all $x, y \in V$.
3. (Identity element of vector addition) There exists an element $0 \in V$, called the **zero element**, such that $0 + x = x + 0 = x$ for every $x \in V$.
4. (Inverse elements of vector addition) For each $x \in V$ there exists an element $-x$, called the **additive inverse** of x, such that $x + (-x) = -x + x = 0$.
5. (Compatibility of scalar multiplication with field multiplication) $a(bx) = (ab)x$ for every $a, b \in \mathbb{R}$ and $x \in V$.
6. (Identity element of scalar multiplication) $1x = x$ for every $x \in V$.
7. (Distributivity of scalar multiplication with respect to vector addition) $a(x+y) = ax+ay$ for every $a \in \mathbb{R}$ and $x, y \in V$.
8. (Distributivity of scalar multiplication with respect to field addition) $(a+b)x = ax+bx$ for every $a, b \in \mathbb{R}$ and $x \in V$.

Example 2.1. For each $d \geq 1$, $\mathbb{R}^d = \{(x_1, \ldots, x_d) : x_1, \ldots, x_d \in \mathbb{R}\}$ is a vector space with operations defined entrywise by

$$(x_1, \ldots, x_d) + (y_1, \ldots, y_d) = (x_1 + y_1, \ldots, x_d + y_d) \quad \text{and} \quad a(x_1, \ldots, x_d) = (ax_1, \ldots, ax_d).$$

Example 2.2. The complex numbers $\mathbb{C}$ are a vector space over $\mathbb{R}$.

Example 2.3. For each $n, m \geq 1$, the space of $n \times m$ matrices is a vector space with operations defined entrywise by

$$(A + B)_{i,j} = A_{i,j} + B_{i,j} \qquad \text{and} \qquad (\lambda A)_{i,j} = \lambda A_{i,j}.$$

(Note that matrix *multiplication* does not feature in the definition of the space of matrices *as a vector space*, which is isomorphic to $\mathbb{R}^{n \times m}$.)

Example 2.4. The space $\mathbb{R}^{[0,1]}$ of *all* functions $[0, 1] \to \mathbb{R}$, the space $C([0, 1]) \to \mathbb{R}$ of *continuous* functions $[0, 1] \to \mathbb{R}$, and the space $C^1([0, 1])$ of *continuously differentiable* functions $[0, 1] \to \mathbb{R}$ are all vector spaces with respect to pointwise addition and scalar multiplication. (It is standard to denote these spaces this way, so that $C^1(\mathbb{R})$ means the space of continuously differentiable functions $\mathbb{R} \to \mathbb{R}$ and so on.)

Example 2.5. Let $I \subseteq \mathbb{R}$ be a non-trivial interval. Given functions $a_0, \ldots, a_{n-1} : I \to \mathbb{R}$, the set of functions

$$\Big\{ f : I \to \mathbb{R} : f^{(n)}(t) + a_{n-1}(t) f^{(n-1)}(t) + \cdots + a_0(t) f(t) = 0 \text{ for every } t \in I \Big\}$$

is a vector space under pointwise addition.

Linear maps. A function $f : V \to W$ between vector spaces is said to be **linear** if $f(ax + by) = af(x) + bf(y)$ for every $a, b \in \mathbb{R}$ and $x, y \in V$. A bijective linear map between vector spaces is called a **linear isomorphism**; the inverse of a linear isomorphism is automatically linear also. We call two vector spaces **isomorphic** if there is a linear isomorphism between them. Given two vector spaces V and W, the space $\mathcal{L}(V, W)$ of linear maps between V and W is itself a vector space under pointwise addition and multiplication, i.e. where we define $(af + bg)(x) = af(x) + bg(x)$ for every $a, b \in \mathbb{R}$, $f, g \in \mathcal{L}(V, W)$ and $x \in V$.

Linear independence, spanning sets and bases. Given a vector space V and a set X of vectors in V, we define the **linear span** of X to be the set of **linear combinations** of elements of X, that is,

$$\langle X \rangle := \big\{ a_1 x_1 + \cdots a_k x_k : k \geq 1,\, x_1, \ldots, x_k \in X,\, a_1, \ldots, a_k \in \mathbb{R} \big\},$$

and say that X **spans** V if $\langle X \rangle = V$. A vector space V is said to be **finite-dimensional** if it admits a finite spanning set. On the other hand, we say that X is **linearly independent** if the equation

$$a_1 x_1 + a_2 x_2 + \cdots a_k x_k = 0$$

holds for some $a_1, \dots a_k \in \mathbb{R}$ if and only if $a_i = 0$ for every $1 \leq i \leq k$. In other words, X is linearly independent if it does not contain zero and $x \notin \langle X \setminus \{x\} \rangle$ for every $x \in X$. A linearly independent set that spans V is called a **basis**[5] of V. For $\mathbb{R}^d$, one choice of basis is given by the vectors

$$\{(1, 0, \dots, 0), (0, 1, 0, \dots, 0), \dots, (0, \dots, 0, 1)\},$$

but this is not at all the only choice of basis!

The following facts are hopefully familiar to you from your previous courses:

Fact 2.6. *A set B is a basis for a vector space V if and only if it is a spanning set that does not have any strict subset that also spans, if and only if it is a linearly independent set that does not have any strict superset that is linearly independent.*

Fact 2.7. *If B_1 and B_2 are both bases of a vector space V then they have the same number of elements. We call this number the **dimension** of V.*

Fact 2.8. *If V_1 and V_2 are vector spaces with the same finite dimension then they are isomorphic. In particular, every d-dimensional vector space is isomorphic to $\mathbb{R}^d$.*

Note that there are many different isomorphisms between two vector spaces of the same dimension: In particular, if we choose a basis for each space then we can always find an isomorphism sending one basis to the other.

Example 2.9. The space $\mathbb{R}^{[0,1]}$ of *all* functions $[0, 1] \to \mathbb{R}$, the space $C([0, 1]) \to \mathbb{R}$ of *continuous* functions $[0, 1] \to \mathbb{R}$, and the space $C^1([0, 1])$ of *continuously differentiable* functions $[0, 1] \to \mathbb{R}$ are all infinite-dimensional spaces. Indeed, the functions $1, x, x^2, \dots$ are an infinite sequence of linearly independent elements in all of these spaces. (These functions do *not* form an algebraic basis.)

Matrices. Every linear map $f : \mathbb{R}^n \to \mathbb{R}^m$ can be represented by an $m \times n$ matrix $A = (A_{i,j})_{1 \leq i \leq n, 1 \leq j \leq m}$ which is the unique array of numbers such that

$$f\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}\right) = \begin{bmatrix} \sum_{j=1}^n A_{1,j} x_j \\ \sum_{j=1}^n A_{2,j} x_j \\ \vdots \\ \sum_{j=1}^n A_{m,j} x_j \end{bmatrix}$$

for every vector $(x_1, \dots, x_n) \in \mathbb{R}^n$. Conversely, given an $n \times m$ matrix A we can define a linear map f using the same relation, which we normally write in shorthand as

$$f(x)_i = \sum_j A_{i,j} x_j.$$

[5] In the infinite-dimensional context this is usually *not* the right notion of basis, and is sometimes called an 'algebraic basis' to distinguish it from more appropriate notions. In this course we will mostly deal with finite dimensional spaces.

Be careful to note that when we write elements of $\mathbb{R}^n$ using coordinates in this manner we are implicitly choosing a basis for our vector space, and that the matrix representation of a linear map depends on this choice of basis!

Composition of linear maps corresponds at the level of matrices to *matrix multiplication*: If A is an $m \times k$ matrix and B is a $k \times n$ matrix then the product AB is an $m \times n$ matrix defined by

$$(AB)_{i,j} = \sum_{\ell=1}^{m} A_{i,\ell} B_{\ell,j}.$$

If A represents the linear map f and B represents the linear map g then AB represents the linear map $f \circ g$.

Since linear maps between finite-dimensional vector spaces can always be uniquely represented as a matrix after choosing a basis for each space, we also have the following fact.

Fact 2.10. *If V and W are finite-dimensional vector spaces of dimension n and m respectively, then $\mathcal{L}(V, W)$ is finite-dimensional with dimension nm.*

2.2 Norms, continuity and differentiability

I next want to review some basic definitions from *multivariable* calculus. Since later on I will want to tell you about matrix-valued ODEs, I want to define what it means for a function *between two finite-dimensional vector spaces* to be (n-times) differentiable. Although you probably haven't encountered this definition before, it is not really any different in content than the definition for functions $\mathbb{R}^n \to \mathbb{R}^m$. It will however require a little more sophistication to state things correctly.

We first have to define what it means for a sequence in a finite-dimensional vector space V to converge. One way to do this is to choose an isomorphism between V and $\mathbb{R}^d$ for some $d \geq 1$, then define convergence coordinatewise. This is fine – one can verify that the resulting notion of convergence does not depend on the choice of isomorphism and that it coincides with the definition we are about to give – but is unsatisfactory in some regards. We will instead work with *norms*. Given a vector space V, a **norm** of V is a function $\|\cdot\| : V \to [0, \infty)$ satisfying

1. $\|x\| = 0$ if and only if $x = 0$[6].
2. (Triangle inequality.) $\|x + y\| \leq \|x\| + \|y\|$ for every $x, y \in V$.
3. $\|\lambda x\| = |\lambda| \cdot \|x\|$ for every $\lambda \in \mathbb{R}$ and $x \in V$.

We think of $\|x\|$ as being the 'distance' from x to the origin as determined by the norm $\|x\|$. (These definitions ensure that $d(x, y) := \|x - y\|$ defines a *metric* on V. Don't worry if you

[6] A function satisfying all the axioms other than this one is called a **seminorm**.

don't know what this means.) A pair $(V, \|\cdot\|)$ where V is a vector space and $\|\cdot\|$ is a norm on V is called a **normed vector space**.

On $\mathbb{R}$, the only norms are of the form $\|x\| = c|x|$ where c is a positive constant. In higher dimensions there are many more norms, with important examples including the ℓ^p **norms** with $1 \leq p \leq \infty$, defined on $\mathbb{R}^d$ by

$$\|(x_1, \ldots, x_d)\|_p = \begin{cases} \left(\sum_{i=1}^d |x_i|^p\right)^{1/p} & p < \infty \\ \max |x_i| & p = \infty. \end{cases}$$

(It is true but not obvious that $\|\cdot\|_p$ always satisfies the triangle inequality when $p \geq 1$ – this is known as *Minkowski's inequality.*) Note that the ℓ^2 norm of $(x_1, \ldots, x_d)$ is just the usual Euclidean distance between x and the origin

$$\|(x_1, \ldots, x_d)\|_2 = \sqrt{\sum_{i=1}^d |x_i|^2}.$$

(NB: We will also use $\|\cdot\|_1, \|\cdot\|_2$ to denote an arbitrary pair of norms. Hopefully it will be clear from context that these are not to be confused with the $p = 1, 2$ cases of the ℓ^p norm.)

Remark 2.11. Analogues of these norms can also be defined on spaces of continuous functions. For example, if $C([0,1], \mathbb{R})$ is the space of continuous functions from $[0,1]$ to $\mathbb{R}$ then

$$\|f\|_p = \begin{cases} \left(\int_0^1 |f(x)|^p\right)^{1/p} & p < 1 \\ \sup_{x \in [0,1]} |f(x)| & p = \infty \end{cases}$$

defines a norm on $C([0,1], \mathbb{R})$ for each $1 \leq p \leq \infty$. The $p = \infty$ norm on $C([0,1], \mathbb{R})$ is also known as the **uniform norm**, and will play an important role throughout the course.

One reason why it makes sense for us to work with general norms is that when we come to study matrices, other choices of norms besides the Euclidean norm in the entries are much more natural. Indeed, if $(V_1, \|\cdot\|_1)$ and $(V_2, \|\cdot\|_2)$ are normed vector spaces, the **operator norm** on $\mathcal{L}(V_1, V_2)$ is defined by

$$\|f\|_{\text{op}} = \sup \left\{ \frac{\|f(x)\|_2}{\|x\|_1} : x \in V \setminus \{0\} \right\}.$$

When $(V_1, \|\cdot\|_1) = (V_2, \|\cdot\|_2)$, the operator norm has the important property that

$$\|f \circ g\|_{\text{op}} \leq \|f\|_{\text{op}} \|g\|_{\text{op}}, \tag{2.1}$$

which is why we will usually want to use operator norms rather than entrywise-defined norms when studying spaces of matrices.

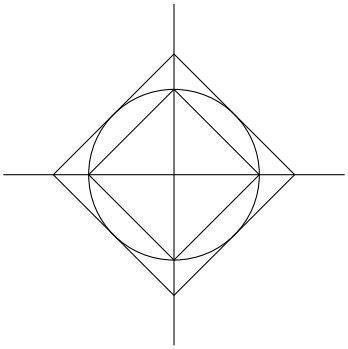


Figure 3: In $\mathbb{R}^2$, the ℓ^2 unit ball contains the ℓ^1 unit ball and is contained in the ℓ^1-ball of radius $\sqrt{2}$.

Exercise 7. Verify that the operator norm is indeed a norm and that it satisfies (2.1). Give an example of linear maps f and g for which (2.1) is strict.

Exercise 8. Prove or provide a counterexample: If $f : V \to V$ is a linear map for some normed space V then $\|f^2\|_{\text{op}} = \|f\|_{\text{op}}^2$.

The following important fact is not very hard to prove, but will be left unproven since it is rather tangential to the rest of the course. You will probably see a full proof in the next analysis course you take.

Fact 2.12. *All norms on a finite-dimensional vector space are "equivalent": If V is finite-dimensional and $\|\cdot\|_1$ and $\|\cdot\|_2$ are two norms on V then there exist positive constants c and C such that*

$$c\|x\|_2 \leq \|x\|_1 \leq C\|x\|_2 \qquad \text{and} \qquad c\|x\|_1 \leq \|x\|_2 \leq C\|x\|_1$$

for every $x \in V$.

This fact has the following geometric interpretation. Given a norm $\|\cdot\|$, the **unit ball** is defined to be $B = \{x \in V : \|x\| \leq 1\}$, and for each $x \in V$ we can write

$$\|x\| = \inf\{\lambda > 0 : x \in \lambda B\}$$

where $\lambda B := \{\lambda x : x \in B\}$. As such, the claim is equivalent to saying that if B_1 and B_2 are the unit balls associated to two different norms on the same finite-dimensional vector space V then if we scale B_2 by a big enough constant it will contain B_1, while if we scale B_2 by a small enough constant it will be contained in B_1 (see Figure 3). This is hopefully intuitively plausible to you, and you might like to think about how to flesh out all the details to turn it into a proof.

Remark 2.13. The analogue of this fact is not true for infinite-dimensional spaces. For example, the different L^p norms on $C([0,1],\mathbb{R})$ are not equivalent to each other. This can be seen

by computing the L^p norm of x^n:

$$\|x^n\|_p = \begin{cases} (1+np)^{-1/p} & p < \infty \\ 1 & p = \infty. \end{cases}$$

If $1 \leq p_1 < p_2 \leq \infty$ then $\|x^n\|_{p_2}/\|x^n\|_{p_1} \to \infty$ as $n \to \infty$, so that the two norms cannot be equivalent.

Remark 2.14. In fact the map sending a norm to its unit ball is a bijection[7] between the set of norms on a finite-dimensional vector space and the set of convex subsets of the space that are non-empty and symmetric (in the sense that $\{-x : x \in B\} = \{x : x \in B\}$ where B is the set). Since there are a lot of convex symmetric sets (in strictly more than one dimension), there are a lot of different norms.

This means that we can define what it means for a sequence to converge in a finite-dimensional vector space using any norm on that space, and the norm we choose will not actually affect the definition[8]: Given a finite-dimensional vector space V and a norm $\|\cdot\|$ on V, we say that a sequence $(x_n)_{n\geq 1}$ in V converges to a point $x \in V$ if $\|x_n - x\| \to 0$ as $n \to \infty$.

Lemma 2.15. *Let V be a finite-dimensional vector space and let $\|\cdot\|$ be a norm on V.*

1. *If $(x_n)_{n\geq 1}$ is a Cauchy sequence in the sense that $\lim_{n\to\infty} \sup_{m\geq n} \|x_m - x_n\| = 0$ then there exists $x \in V$ such that $\|x_n - x\| \to 0$ as $n \to \infty$.*

2. *If $(x_n)_{n\geq 1}$ is a sequence in V such that $\sum_{n=1}^{\infty} \|x_n\| < \infty$ then $\sum_{n=1}^{\infty} x_n$ is well-defined in the sense that the partial sums $\sum_{n=1}^{N} x_n$ converge to some element of V as $N \to \infty$.*

Proof. The first part is true for $\mathbb{R}^d$ and is therefore true for any finite-dimensional normed vector space since they are all isomorphic and all norms are equivalent. The second part follows since if $\sum_{n=1}^{\infty} \|x_n\| < \infty$ then the sequence of partial sums $(\sum_{n=1}^{N} x_n)$ is Cauchy since

$$\Big\|\sum_{n=1}^{N} x_n - \sum_{n=1}^{M} x_n\Big\| = \Big\|\sum_{n=\min\{N,M\}}^{\max\{N,M\}} x_n\Big\| \leq \sum_{n=\min\{N,M\}}^{\max\{N,M\}} \|x_n\| \leq \sum_{n=\min\{N,M\}}^{\infty} \|x_n\|,$$

which is small if $\min\{N, M\}$ is large since $\sum \|x_n\|$ converges. □

Similar considerations allow us to define continuity, differentiability etc of maps between (subsets of) finite-dimensional vector spaces by first choosing norms on these spaces, with the resulting definitions not depending on the choices we make by Fact 2.12. Let V_1 and V_2 be

[7] Recall that a function $f : X \to Y$ between two sets is said to be **injective** if $f(x_1) \neq f(x_2)$ for every pair of distinct elements $x_1, x_2 \in X$, **surjective** if for every $y \in Y$ there exists $x \in X$ such that $f(x) = y$, and **bijective** if it is both injective and surjective. Each bijective function f has a unique **inverse** $f^{-1} : Y \to X$ satisfying $f^{-1}(f(x)) = x$ for every $x \in X$ and $f(f^{-1}(y)) = y$ for every $y \in Y$.

[8] In infinite-dimensional spaces (such as function spaces) the distinction between different norms is much more important, something you will learn about in detail if you take a course in functional analysis.

finite-dimensional vector spaces and let $\|\cdot\|_1$ and $\|\cdot\|_2$ be norms on V_1 and V_2 respectively. Given a subset $\Omega \subseteq V_1$ and a function $f : \Omega \to V_2$, we say that f is **continuous** at a point $x \in \Omega$ if for every $\varepsilon > 0$ there exists $\delta > 0$ such that if $y \in \Omega$ satisfies $\|x - y\|_1 \leq \delta$ then $\|f(x) - f(y)\| \leq \varepsilon$, and say that f is continuous if it is continuous at every point of Ω.

We say that f is **differentiable** at $x \in \Omega$ if there exists a linear map $Df(x) : V_1 \to V_2$ such that

$$\frac{\|f(y) - f(x) - Df(x)(y - x)\|_2}{\|y - x\|_1} \to 0 \qquad \text{as } y \to x \text{ in } \Omega,$$

and say that f is differentiable if it is differentiable at every point of Ω, in which case Df defines a function $Df : \Omega \to \mathcal{L}(V_1, V_2)$. Since $\mathcal{L}(V_1, V_2)$ is itself a finite-dimensional vector space, we also have a well-defined notion of what it means for Df to be continuous or differentiable, leading iteratively to a definition of what it means for a function $f : \Omega \to V_2$ to be n-times differentiable. We say that f is **smooth** if it is n-times differentiable for every $n \geq 1$.

Exercise 9. Using fact 2.12, check carefully that the definitions of continuity, differentiability, and smoothness do not depend on the choice of norm when the relevant vector spaces are finite-dimensional.

Exercise 10. Prove that linear functions between finite-dimensional spaces are always smooth.

Note that there is an interesting distinction here between the one-dimensional and higher-dimensional cases: For smooth vector-valued functions $f : \mathbb{R} \to V$, all of the derivatives $f^{(n)}$ can also be thought of as functions $\mathbb{R} \to V$ since there is a canonical isomorphism $\mathcal{L}(\mathbb{R}, V) \equiv V$ given by $f \mapsto f(1)$. In contrast, for a smooth function $f : \mathbb{R}^a \to \mathbb{R}^b$, the nth deriative $D^n f$ is a function $\mathbb{R}^a \to \mathcal{L}(\mathbb{R}^a, \mathcal{L}(\cdots \mathcal{L}(\mathbb{R}^a, \mathcal{L}(\mathbb{R}^a, \mathbb{R}^b)) \cdots))$, where the space on the right has dimension $a^n b$. As such, even for functions $f : \mathbb{R}^2 \to \mathbb{R}$ the higher derivatives of f can be very complicated, high-dimensional objects. From now on we will always make the identification $\mathcal{L}(\mathbb{R}, V) \equiv V$ and consider all derivatives of a function $f : \mathbb{R} \to V$ as taking values in V.

Exercise 11. (Optional) Verify that the usual rules of calculus extend to functions between finite-dimensional vector spaces. (Here, given functions into spaces of linear maps, I am writing $\circ$ to denote composition *of the functions* and $\cdot$ to denote composition *of the linear maps*. This is not standard.)

1. (Linearity) Let V_1 and V_2 be finite-dimensional vector spaces and let Ω be a subset of V_1. Prove that if $f, g : \Omega \to V_2$ are differentiable then $af + bg$ is differentiable with derivative $D(af + bg) = aDf + bDg$ for every $a, b \in \mathbb{R}$.

2. (Chain rule) Let V_1, V_2, and V_3 be finite-dimensional vector spaces and let Ω_1 and Ω_2 be subsets of V_1 and V_2 respectively. Prove that if $f : \Omega_1 \to \Omega_2$ and $g : \Omega_2 \to V_3$ are differentiable then $g \circ f$ is differentiable with

$$[D(g \circ f)](x) = [Dg](f(x)) \cdot [Df](x)$$

for every $x \in \Omega_1$.

3. (Product rule) Let V_1, V_2, V_3 and V_4 be finite-dimensional vector spaces and let Ω be a subsets of V_1. Prove that if $f : \Omega \to \mathcal{L}(V_2, V_3)$ and $g : \Omega \to \mathcal{L}(V_3, V_4)$ are differentiable then $g \cdot f$ is differentiable with

$$[D(g \cdot f)](x) = [Dg](x) \cdot f(x) + g(x) \cdot [Df](x)$$

4. Let V_1, V_2, and V_3 be finite-dimensional vector spaces, let $n \geq 1$, and let Ω_1 and Ω_2 be subsets of V_1 and V_2 respectively. Prove that if $f : \Omega_1 \to \Omega_2$ and $g : \Omega_2 \to V_3$ are n-times differentiable then $g \circ f$ is n-times differentiable. (Hint: Induct on n using parts 1-3.)

Open sets. We now discuss the (very important!) notion of an **open set** in a normed vector space: Given a normed vector space $(V, \|\cdot\|)$, we say that a set $U \subseteq V$ is **open** if for each $x \in U$ there exists $\varepsilon > 0$ such that $\{y \in V : \|y - x\| \leq \varepsilon\} \subseteq U$. Note that if two norms on the same space are equivalent in the sense of fact 2.12 then they have the same open sets. In particular, the open sets of a finite-dimensional vector space are defined independently of the choice of norm.

Some example are in order:

1. If V is a vector space then V and the empty set $\emptyset$ are open in V with respect to any norm.

2. An open interval (a, b) with $a < b$ is open in $\mathbb{R}$, as are the half-infinite open intervals $(-\infty, b)$ and (a, ∞).

3. Open rectangles $\prod_{i=1}^{d}(a_i, b_i)$ in $\mathbb{R}^d$ are open in $\mathbb{R}^d$.

4. Closed rectangles $\prod_{i=1}^{d}[a_i, b_i]$ in $\mathbb{R}^d$ are *not* open. For example, the interval $[0, 1]$ is not open in $\mathbb{R}$ since it does not contain any set of the form $(-\varepsilon, \varepsilon)$ or $(1 - \varepsilon, 1 + \varepsilon)$.

5. If V is a finite-dimensional vector space and $A \subseteq V$ is a finite set of points then $V \setminus A = \{x \in V : x \notin A\}$ is open in V.

Exercise 12. Prove that a function $f : \mathbb{R}^n \to \mathbb{R}^m$ is continuous if and only if for each open set $U \subseteq \mathbb{R}^m$, the set $f^{-1}(U) = \{x \in \mathbb{R}^n : f(x) \in U\}$ is open in $\mathbb{R}^n$.

Integration in vector spaces. It will be useful to know how to *integrate* vector-valued functions as well as differentiate them. Suppose that $f : I \to V$ is a continuous function from a non-trivial interval $I \subseteq \mathbb{R}$ to a finite-dimensional vector space V. For each pair of numbers $a < b$ in I, the integral $\int_a^b f(t)\,\mathrm{d}t$ is defined to be the unique vector in V such that if $\{e_1, \ldots, e_n\}$ is a basis for V and we represent f in terms of these basis elements as

$$f(t) = \sum_{i=1}^{n} f_i(t) e_i,$$

where each of the function $f_i : I \to \mathbb{R}$ is continuous (as justified in the exercise below), then

$$\int_a^b f(t)\,\mathrm{d}t := \sum_{i=1}^n \left(\int_a^b f_i(t)\,\mathrm{d}t \right) e_i.$$

Exercise 13. Let V be a finite-dimensional vector space and let $f : I \to \mathbb{R}$ be a function from a non-trivial interval to V. Let $\{e_1, \ldots, e_n\}$ be a basis of V and let

$$f(t) = \sum_{i=1}^n f_i(t) e_i$$

for each $t \in I$. Prove that f is continuous if and only if f_i is continuous for every $1 \leq i \leq n$.

Exercise 14. Prove that the definition of $\int_a^b f(t)\,\mathrm{d}t$ does not depend on the choice of basis.

The fundamental theorem of calculus extends to this setting as follows:

Exercise 15. Let V be a finite-dimensional vector space and let $f : I \to \mathbb{R}$ be a continuous function from a non-trivial interval to V. Prove that if $t_0 \in I$ then $F(t) = \int_{t_0}^t f(s)\,\mathrm{d}s$ is differentiable with derivative f.

2.3 ODEs in vector spaces

At the beginning of the course we considered only ODEs whose outputs were one-dimensional. For many applications, we will want to consider the higher-dimensional case also. As we will explain in detail in the next section, one reason to do this is to work in phase space and take all ODEs to be first-order. Of course there are also many examples where the quantities of interest are already multi-dimensional before we pass to phase space by also tracking their derivatives, such as particles moving in three-dimensional space; some specific examples are given at the end of this section.

Let us now give some formal definitions that will be used throughout the rest of the course. Let V be a finite-dimensional vector space. (As usual, we identify $\mathcal{L}(\mathbb{R}, V)$ with V so that derivatives of functions from (intervals in) $\mathbb{R}$ to V are also considered to take values in V.) Given $n \geq 1$, an n**th ODE in** V is an equation[9] of the form

$$f^{(n)} = F(t, f, \ldots, f^{(n-1)}), \tag{ODE}$$

where U where U is an open subset of $\mathbb{R} \times V^n$ and $F : U \to \mathbb{R} \times V^n$ is a function. (Often U will be the whole space. Sometimes our ODE is not defined on the whole space and we need

[9] If you've read too much set theory and have begun to doubt that "'equations" are valid mathematical objects (I don't endorse this opinion), you can consider the ODE to be the data (U, F). You could also define the "equation" to be the set of all its solutions, among various other options.

to define it on a smaller set instead.) An nth order **initial value problem** (IVP) in V is a system of equations of the form

$$(f(t_0), \dots, f^{(n-1)}(t_0)) = \mathbf{x}_0 \qquad \text{and} \qquad f^{(n)} = F(t, f, \dots, f^{(n-1)}) \tag{IVP}$$

where $F : U \to \mathbb{R} \times V^n$ is as above, $t_0 \in \mathbb{R}$ and $\mathbf{x}_0 \in V^n$.

Remark 2.16. Usually we won't be quite this formal and just write our ODE in the natural way, where we take U to be the largest set where the ODE is defined; this natural domain of definition will typically be open when the function F is sufficiently continuous. I will only draw further attention to this if there's something subtle going on and we need to be concerned about it.

Given a pair (I, f) consisting of a non-trivial interval $I \subseteq \mathbb{R}$ and an n-times differentiable function $f : I \to \mathbb{R}^d$, we say that (I, f) is a **solution** to (ODE) if $(t, f, \dots, f^{(n-1)}) \in U$ and $f^{(n)} \equiv F(t, f, \dots, f^{(n-1)})$ for every $t \in I$. We say that (I, f) is a solution to the initial value problem (IVP) if it is a solution to the ODE, $t_0 \in I$, and $(f(t_0), \dots, f^{(n-1)}(t_0)) = \mathbf{x}_0$. We will also use this terminology when studying solutions to ODEs taking values in one dimension, where $V = \mathbb{R}$.

Let us now mention a few examples of higher-dimensional ODEs arising in practice.

Point charges in magnetic fields

Consider a massive point-particle with mass m and charge q (in some fixed unit system, let's say SI) moving through a constant electric field $\mathbf{E}$ and magnetic field $\mathbf{B}$. The equations of motion of the particle can be written

$$\mathbf{a} = \frac{q}{m}(\mathbf{E} + \mathbf{v} \times \mathbf{B}) - \mathbf{g},$$

where $\mathbf{a}$ is the acceleration of the particle, $\mathbf{v}$ is its velocity, and $-\mathbf{g}$ is the downward-pointing vector representing acceleration due to gravity. We can either think of this as a second-order ODE describing the position of the particle or as a first-order ODE describing the velocity; in either case the output at each time is naturally a three-dimensional vector. This is a constant coefficient linear ODE, and you will easily be able to solve it exactly by the end of this section.

Geodesics on a surface

Suppose we have a surface $S \subseteq \mathbb{R}^3$ defined as the zero-set of some smooth function $f(x, y, z) = 0$. A **geodesic** in S is a locally-length-minimizing curve in S. That is, it is a curve in S so that for any two points on the curve that are sufficiently close to one another, there is no strictly shorter curve connecting the same two points. Examples include straight lines in $\mathbb{R}^2$ and great circles in the sphere. If we take a point on S and a unit vector tangent to S at that point, there is a unique geodesic emanating from the point with initial direction given by this unit vector; this geodesic can be thought of as a solution to a certain second-order IVP in $\mathbb{R}^3$ called the **geodesic equation**, where the details of the equation depend on the choice of surface. (For straight lines in $\mathbb{R}^2$ this ODE is just $f'' \equiv 0$.)

Rabbits and wolves: The Lotka-Volterra equations

The **Lotka-Volterra equation** is a non-linear, first-order, two-dimensional ODE used to describe the dynamics of biological systems in which two species interact:

$$\frac{d}{dt}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \alpha x - \beta xy \\ \delta xy - \gamma y \end{pmatrix}.$$

Here, x and y represent the populations of prey (e.g. rabbits) and predators (e.g. wolves), respectively. The prey population is assumed to have an unlimited food supply and to reproduce exponentially in the absence of predators. The predator population has sufficient food supply from the prey and will starve or leave if the prey is extinct. The parameters α, β, γ, and δ are positive constants that represent the reproduction rate of prey, the mortality rate of prey due to predation, the natural death rate of predators, and the reproduction rate of predators due to predation, respectively. It turns out that this nonlinear two-dimensional ODE has *periodic* solutions, but that (in general) these solutions do not admit exact formulas in terms of "standard" functions.

Remark 2.17. Higher-dimensional ODEs such as those we have discussed here are often written as *systems of ODEs*, so that e.g. one might write the Lotka-Volterra equations as

$$x' = \alpha x - \beta xy \qquad \text{and} \qquad y' = \delta xy - \gamma y.$$

The two formulations are completely equivalent. Writing the equation as a single higher-dimensional equation may seem needlessly sophisticated, but it is often much nicer to work with once you are used to it.

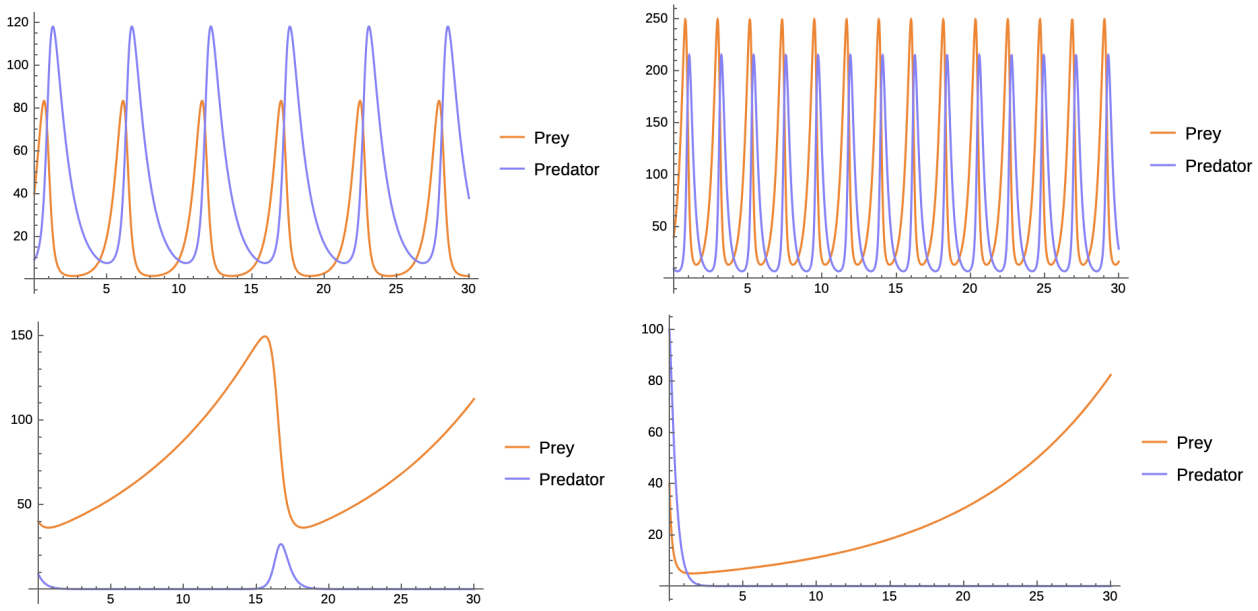


Figure 4: Solutions to the Lotka-Volterra equation with various different parameter values and initial conditions.

2.4 Phase space and reduction to first order

We now discuss how we can always think of all ODEs as being *first order* provided that we are willing to work in more than one dimension. We do this by moving to a bigger space in which we keep track of everything that is relevant for the evolution of our system. This is called "working in **phase space**": The phase space of a general nth order ODE is the space of all possible times, positions, and first $n-1$ derivatives of the position.

Let's first see how this works in a simple example.

Example 2.18. Let (I, f) be a solution to the second-order ODE $f'' = -f$. By considering the function $\mathbf{f} : I \to \mathbb{R}^2$ defined by $\mathbf{f}(t) = (f(t), f'(t))$, we can rewrite the ODE as a first-order ODE in phase space:

$$\mathbf{f}'(t) = \frac{d}{dt}\begin{pmatrix} f \\ f' \end{pmatrix} = \begin{pmatrix} f' \\ f'' \end{pmatrix} = \begin{pmatrix} f' \\ -f \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \mathbf{f}(t).$$

In fact, (I, f) is a solution to the second-order ODE $f'' = -f$ *if and only if* $(I, \mathbf{f})$ is a solution to the first-order ODE $\mathbf{f}'(t) = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right) \mathbf{f}(t)$, where $\mathbf{f}(t) = (f, f')$, *if and only if* there exists a solution $(I, \mathbf{g})$ to this first-order ODE such that f is equal to the first coordinate of $\mathbf{g}$; the fact that the second coordinate of $\mathbf{g}$ is equal to the derivative of the first coordinate is an automatic consequence of the ODE. The fact that we get a *rotation matrix* in this phase-space ODE is closely related to the fact that we get trig functions as solutions to the original second-order ODE. (We will revisit this example in some more detail later.)

Proposition 2.19 (Every ODE can be thought of as a first-order ODE in a higher dimension)**.** *Let $n, d \geq 1$, let $U \subseteq \mathbb{R}^{1+nd}$ be an open set, and let F be a function $F : U \to \mathbb{R}^d$. Define a function $\mathbf{F} : U \to \mathbb{R}^{nd}$ by*

$$\mathbf{F}(t, x_0, x_1, \ldots, x_{n-1}) = \big(x_1, x_2, \ldots, x_{n-1}, F(t, x_0, \ldots, x_{n-1})\big)$$

for every $(t, x_0, \ldots, x_{n-1}) \in U$. Then

$$(I, f) \mapsto (I, \mathbf{f}) \qquad \textit{where} \qquad \mathbf{f} := (f, f', \ldots, f^{(n-1)})$$

defines a bijection

$$\{\textit{solutions of } f^{(n)} = F(t, f, \ldots, f^{(n-1)})\} \to \{\textit{solutions of } \mathbf{f}' = \mathbf{F}(t, \mathbf{f})\}.$$

The inverse of this bijection is defined by taking $(I, \mathbf{f}) \to (I, \mathbf{f}_0)$, where $\mathbf{f} = (\mathbf{f}_0, \ldots, \mathbf{f}_{n-1})$.

Remark 2.20. Note that when $d > 1$, the coordinates of $\mathbf{f}$ are themselves more than one-dimensional. If you find this confusing, it may help you to think that

$$\mathbf{f} = ((f_1, \ldots, f_d), (f_1', \ldots, f_d'), \ldots, (f_1^{(n-1)}, \ldots, f_d^{(n-1)})).$$

While this may suggest we should use different notation for points in the canonically isomorphic spaces $\mathbb{R}^{dn}$ and $(\mathbb{R}^d)^n$, we will generally avoid doing so in this course.

For our purposes, this proposition means that we can often restrict attention to the first-order case when proving theorems about ODEs. It does *not* mean that it is always best to view specific higher-order ODEs through the lens of this reduction when working with examples, as there are also significant advantages to working with functions with one-dimensional range.

Proof. The reasoning is no different than in the simple example we saw above. We can think of the equation $\mathbf{f}' = \mathbf{F}(t, \mathbf{f})$ governing the vector-of-vectors $\mathbf{f} = (\mathbf{f}_0, \ldots, \mathbf{f}_{n-1})$ as a system of equations

$$\begin{aligned} \mathbf{f}_i' &= \mathbf{f}_{i+1} && i = 0, \ldots, n-2 \\ \mathbf{f}_{n-1}' &= F\big(t; \mathbf{f}_0(t), \mathbf{f}_1(t), \mathbf{f}_2(t), \ldots, \mathbf{f}_{n-1}(t)\big)\,. \end{aligned}$$

Solving the first part of this system is equivalent to $\mathbf{f}_i$ being the ith derivative of $\mathbf{f}_0$ for each $1 \leq i \leq n-1$, so that $\mathbf{f}$ solves the entire system if and only if $\mathbf{f}_i$ is the ith derivative of $\mathbf{f}_0$ for each $1 \leq i \leq n-1$ and $f = \mathbf{f}_0$ satisfies the ODE $f^{(n)} = F(t, f, f', \ldots, f^{(n-1)})$. □

Phase spaces in mechanics

For a ball in flight we could take the phase space to simply be the position and velocity of the ball; more accurate treatment might include additional variables to keep track of how the ball is spinning etc., so that phase space for the dynamics of a simple object can be reasonably high-dimensional. If we considered the time evolution of a system of n point-projectiles (without considering things like spin), the "obvious" way of writing down the system would be as a $3n$-dimensional second-order ODE, while the phase space representation would be a first order ODE in $6n$ dimensions.

A further remark: Since the ODEs arising in classical mechanics are usually second-order, the relevant phase spaces usually have an *even* number of dimensions, regardless of how many degrees of freedom we are keeping track of in the system. *Symplectic geometry* is an area of mathematics studying the abstract structure of even-dimensional spaces (equipped with a notion of time-evolution) of the kind arising as phase-spaces in classical mechanics.

Tangent bundles

Here we continue our discussion of geodesics on a surface from the last section. The relevant phase space here is called the 'unit tangent bundle', and consists of all pairs of a point in the surface and a unit tangent vector emanating from that point; the unit tangent bundle of a surface in $\mathbb{R}^3$ can be thought of as a "three-dimensional object" (technically speaking, a 3-dimensional manifold) living in $\mathbb{R}^6$. This space is not the full phase-space of all solutions to the ODE, it is just a subspace describing the geodesics inside the surface that are parameterised by arc-length. (The full tangent bundle, where tangent vectors are not required to be unit vectors, is a *four-dimensional* object in $\mathbb{R}^6$.) "Bundles" like this also arise naturally when one wants to define what it means to do calculus or study ODEs "inside a surface" (or any other curved space) and are very important both in mathematical subjects like geometry and topology and in general relativity.

Exercise 16. Formulate and prove a version of Proposition 2.19 stating that every ODE can be thought of as an *autonomous* first-order ODE in a higher dimension, i.e., an ODE of the form $\mathbf{f}' = F(\mathbf{f})$, where there is no dependence on time on the right hand side. (Note: You will want to use a different vector $\mathbf{f}$ than we used above.)

2.5 Constant coefficient linear ODEs and matrix exponentiation

Let $U \subseteq \mathbb{R}$ be open and consider the nth order linear ODE defined by

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

where $a_0, \ldots, a_{n-1}, b : U \to \mathbb{R}$. As usual, we can always think of an nth order ODE as a first order ODE in phase space, which we can write in vector notation as

$$\begin{aligned} \frac{d}{dt}\begin{pmatrix} f^{(n-1)} \\ f^{(n-2)} \\ \vdots \\ f' \\ f \end{pmatrix} &= \begin{pmatrix} b - a_{n-1}f^{(n-1)} - \cdots - a_0 f \\ f^{(n-1)} \\ \vdots \\ f'' \\ f' \end{pmatrix} \\ &= \begin{pmatrix} b \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} -a_{n-1} & -a_{n-2} & \cdots & -a_2 & -a_1 & -a_0 \\ 1 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 \\ 0 & 0 & \cdots & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} f^{(n-1)} \\ f^{(n-2)} \\ \vdots \\ f' \\ f \end{pmatrix}. \end{aligned} \tag{2.2}$$

If we write $\mathbf{f} = (f^{(n-1)}, \ldots, f)$, $\mathbf{b} = (b, \ldots, 0)$, and write A for the $n \times n$ matrix given by $A_{1,j} = -a_{n-j}$ and $A_{i,j} = \mathbb{1}(j = i-1)$ for each $i \neq 1$, we can therefore express our linear ODE in the very convenient form

$$\mathbf{f}'(t) = \mathbf{b}(t) + A(t)\mathbf{f}(t).$$

In particular, if the equation is *homogeneous*, meaning that $b \equiv 0$, we have the even simpler form

$$\mathbf{f}' = A(t)\mathbf{f}(t).$$

We will focus on the homogeneous case for now and return to the inhomogeneous case later.

We now give an informal overview of how we will proceed; we will fill the details in later. We know that if we have a homogeneous linear ODE with two solutions f_1 and f_2 then any linear combination of f_1 and f_2 is also a solution. It follows that if the a_i are continuously differentiable, I is a non-trivial interval, and $t_0 \in I$ is such that there is a solution (I, f) to the IVP $f^{(n)} + a_{n-1}f^{(n-1)} + \cdots + a_1 f' + a_0 f = 0$, $\mathbf{f} = \mathbf{x}_0$ for every $\mathbf{x}_0 \in \mathbb{R}^n$ then, for each $t \in I$, the map sending the initial condition $\mathbf{x}_0$ to the value of the solution to the relevant IVP at t is linear! In other words, in this situation there will be a *matrix valued function* M such that $\mathbf{f}(t) = M(t)\mathbf{f}(t_0)$ for every t, and this matrix-valued function must satisfy

$$(M(t)\mathbf{f}(t_0))' = A(t)M(t)\mathbf{f}(t_0).$$

Assuming for now that the usual rules of calculus apply to matrix-valued functions, this

should be equivalent to M satisfying the ODE

$$M'(t) = A(t)M(t).$$

At least superficially, this looks just like the one-dimensional ODE

$$f' = gf$$

which we saw had the solution

$$f = \exp\left[\int g(s)\,\mathrm{d}s\right].$$

It's therefore reasonable to hope that we can solve our matrix ODE in a similar way, so that

$$M(t) = \exp\left[\int A(s)\,\mathrm{d}s\right]$$

and we will see that this can be done *provided that the matrices $A(t)$ all commute with each other*. Of course, a first step towards doing this should be to figure out what it means to exponentiate a matrix!

Calculus with matrices and matrix exponentiation. We now apply the general setup described in the previous section to do *calculus with matrices*. Let $\mathcal{M}(m,n)$ be the space of $n \times m$ **complex** matrices, which describe **complex linear maps** $\mathbb{C}^n \to \mathbb{C}^m$. We also write $\mathcal{M}(n) = \mathcal{M}(n,n)$. While this vector space has dimension nm as a complex vector space, we will instead think of it as a *real* vector space of dimension $4nm$. We will always equip $\mathbb{C}^n$ with its usual Euclidean norm

$$\|(x_1 + iy_1, \ldots, x_n + iy_n)\| = \sqrt{\sum_{i=1}^{n}(x_i^2 + y_i^2)}$$

and equip $\mathcal{M}(m,n)$ with the associated operator norm $\|\cdot\|_{\mathrm{op}}$.

We already know how to differentiate and integrate functions from (intervals in) $\mathbb{R}$ to $\mathcal{M}(m,n)$ as a special case of our general constructions concerning finite-dimensional vector spaces. To really do 'calculus with matrices' rather than just 'calculus in vector spaces' we need the matrix version of the product rule.

Lemma 2.21. *Let n, m, k be positive integers, let I be a non-trivial interval, and suppose that $A : I \to \mathcal{M}(m,k)$ and $B : I \to \mathcal{M}(k,n)$ are differentiable. Then their product AB is differentiable with derivative $(AB)' = A'B + AB'$.*

Proof. Apply the usual product rule to each entry $(AB)_{i,j}(t) = \sum_\ell A_{i,\ell}(t)B_{\ell,j}(t)$ to get that

$$(AB)'_{i,j} = \sum_\ell A'_{i,\ell}B_{\ell,j} + A_{i,\ell}B'_{\ell,j} = (A'B)_{i,j} + (AB')_{i,j}$$

for every $1 \leq i \leq m$ and $1 \leq j \leq n$. $\square$

We next define matrix exponentiation by mimicking the usual exponentiation of numbers. Given a square matrix A, we want to define a matrix-valued function e^{tA} that such that

1. e^{0A} is the identity matrix, and
2. $\frac{d}{dt}e^{tA} = Ae^{tA}$ for every $t \in \mathbb{R}$.

Rather than immediately writing down the right definition, let us see how we could arrive at it heuristically. Fix $t > 0$ and suppose that k is a very large integer. Since we want our matrix eponential to satisfy $\frac{d}{dt}e^{tA} = Ae^{tA}$, we should have that

$$e^{tA} = e^{\frac{k-1}{k}tA} + \int_{\frac{k-1}{k}t}^{t} Ae^{sA}\,\mathrm{d}s \approx \left(1 + \frac{t}{k}A\right) e^{\frac{k-1}{k}tA},$$

where the (non-rigorous) estimate in the second line is saying that e^{sA} does not change much as s varies between $\frac{k-1}{k}t$ and t. Applying the same estimates recursively yields (non-rigorously) that

$$e^{tA} \approx \left(1 + \frac{t}{k}A\right)^2 e^{\frac{k-2}{k}tA} \approx \cdots \approx \left(1 + \frac{t}{k}A\right)^k$$

and we can expand this estimate using the binomial theorem

$$e^{tA} \approx \sum_{i=0}^{k} \binom{k}{i} \left(\frac{t}{k}\right)^i A^i.$$

If we write the binomial coefficient as the ratio of $k(k-1)\cdots(k-i)$ to $i!$ and note that $k(k-1)\cdots(k-i)/k^i \to 1$ as $k \to \infty$ for each fixed i, assume that we can safely switch the order of summation and limits, and assume that taking the limit as $k \to \infty$ gives an *exact* expression for e^{tA}, we get that

$$e^{tA} = \sum_{i=0}^{\infty} \frac{(tA)^i}{i!}.$$

In other words, this non-rigorous calculation suggests that we should define matrix exponentials by just 'plugging a matrix in' to the usual power series representation of exponentials of numbers. Of course if we want to use this idea rigorously we should check that doing this a) makes sense, and b) satisfies the properties we would like an exponential to have.

Lemma 2.22. *Let $A \in \mathcal{M}(n)$ be a complex $n \times n$ matrix. The matrix exponential defined by*

$$e^{tA} = \sum_{n=0}^{\infty} \frac{(tA)^n}{n!}$$

is well-defined for every $t \in \mathbb{R}$ in the sense that the infinite sum on the right hand side converges for every $t \in \mathbb{R}$.

Proof. Since $\|(tA)^i\|_{\mathrm{op}} \leq |t|^i \cdot \|A\|_{\mathrm{op}}^i$ for every $n \geq 1$ and $t \in \mathbb{R}$, we have that

$$\sum_{i=0}^{\infty} \left\| \frac{(tA)^i}{i!} \right\|_{\mathrm{op}} \leq \sum_{i=0}^{\infty} \frac{|t|^i \|A\|_{\mathrm{op}}^n}{i!} = e^{|t| \cdot \|A\|_{\mathrm{op}}}$$

for every $t \in \mathbb{R}$. It follows that the series $\sum_{i=0}^{\infty} \frac{(tA)^n}{i!}$ converges to a matrix that has operator norm at most $e^{|t| \cdot \|A\|_{\mathrm{op}}}$ for every $t \in \mathbb{R}$. □

Lemma 2.23. *Let $A \in \mathcal{M}(n)$ be a complex $n \times n$ matrix. Then the matrix exponential e^{tA} commutes with every matrix that A commutes with for every $t \in \mathbb{R}$. In particular, $Ae^{tA} = e^{tA}A$.*

Proof. We will prove more generally that if $(B_k)_{k \geq 1}$ is a sequence of matrices in $\mathcal{M}(n)$ that all commute with some fixed matrix $A \in \mathcal{M}(n)$ and that converge to some matrix B then A and B commute: The claim will follow by applying this to the sequence of partial sums $\sum_{i=0}^{k} \frac{(tA)^i}{i!}$, which all commute with every matrix that A commutes with.

To prove this more general statement, we simply observe that

$$\begin{aligned} \|BA - AB\|_{\mathrm{op}} &\leq \|BA - B_kA\|_{\mathrm{op}} + \|B_kA - AB_k\|_{\mathrm{op}} + \|AB_k - AB\|_{\mathrm{op}} \\ &\leq \|B - B_k\|_{\mathrm{op}}\|A\|_{\mathrm{op}} + 0 + \|A\|_{\mathrm{op}}\|B - B_k\|_{\mathrm{op}} = 2\|A\|_{\mathrm{op}}\|B - B_k\|_{\mathrm{op}} \end{aligned}$$

for every $k \geq 1$, and since the right hand side tends to zero as $k \to \infty$ it follows that $\|BA - AB\|_{\mathrm{op}} = 0$ and hence that $AB = BA$ as claimed. □

Lemma 2.24. *Let $A \in \mathcal{M}(n)$ be a complex $n \times n$ matrix. The matrix exponential e^{tA} defines a smooth function $\mathbb{R} \to \mathcal{M}(n)$ with kth derivative $(e^{tA})^{(k)} = A^k e^{tA}$ for every $k \geq 1$.*

We will deduce this lemma from some general theory that we will prove later in the course. If $I = [a, b]$ is a non-trivial closed bounded interval (bounded means a and b are finite) and $(V, \|\cdot\|)$ is a normed vector space we write $C(I, V)$ for the vector space of continuous functions $f : I \to V$ equipped with the norm

$$\|f\|_\infty := \sup_{t \in I} \|f(t)\|.$$

Theorem 2.25 (Differentiating term-by-term)**.** *Let I be a non-trivial closed bounded interval, let $(V, \|\cdot\|)$ be a finite-dimensional normed vector space, and let $(f_i)_{i \geq 1}$ be a sequence of functions from I to V. Suppose that $n \geq 1$ is such that f_i is n-times differentiable for every $i \geq 1$, the series $\sum_{i=1}^{\infty} f_i^{(m)}(t)$ converges in V for each $t \in I$, and*

$$\sum_{i=1}^{\infty} \|f_i^{(n)}\|_\infty < \infty.$$

Then $\sum_{i=1}^{\infty} f_i$ *is* n*-times differentiable with*

$$\frac{d^m}{dt^m} \sum_{i=1}^{\infty} f_i(t) = \sum_{i=1}^{\infty} f_i^{(m)}(t)$$

for every $1 \leq m \leq n$ *and* $t \in I$*.*

We stress that we only require *pointwise* convergence for the first $n-1$ derivatives, but require the stronger condition of *uniform* convergence for the final derivative. We will prove this theorem later in the course, after we finish our discussion of linear ODEs. Let us now see how it implies Lemma 2.24.

Proof of Lemma 2.24. We can compute the kth derivative of the partial sums defining the matrix exponential to be

$$\sum_{i=0}^{N} \frac{d^k}{dt^k} \frac{t^i A^i}{i!} = \sum_{i=k}^{N} \frac{t^{i-k} A^i}{(i-k)!} = A^k \sum_{i=0}^{N-k} \frac{t^i A^i}{i!}.$$

The claim follows from Theorem 2.25 since if I is any closed bounded interval then

$$\sum_{i=0}^{\infty} \left\| \frac{d^k}{dt^k} \frac{t^i A^i}{i!} \right\|_{\infty} \leq \sum_{i=0}^{\infty} \frac{1}{i!} \left(\|A\|_{\text{op}} \cdot \sup_{t \in I} |t| \right)^i < \infty,$$

and since a function whose restriction to any closed bounded interval is smooth is smooth everywhere. □

Lemma 2.26 (Semigroup property)**.** *Let* $A \in \mathcal{M}(n)$*. Then* $e^{tA} e^{sA} = e^{sA} e^{tA} = e^{(t+s)A}$ *for every* $s, t \in \mathbb{R}$*. In particular,* e^{-tA} *is the inverse of* e^{tA} *for every* $t \in \mathbb{R}$*.*

Proof. It suffices to prove that $e^{tA} e^{sA} = e^{(t+s)A}$; the other equality follows by symmetry. We first claim that e^{-tA} is the inverse of e^{tA} for each $t \in \mathbb{R}$. This is obvious when $t = 0$, in which case both matrices are the identity. We can use the product rule together with Lemma 2.24 to compute that

$$(e^{-tA} e^{tA})' = -Ae^{-tA} e^{tA} + e^{-tA} Ae^{tA} = -Ae^{-tA} e^{tA} + Ae^{-tA} e^{tA} = 0,$$

where we used Lemma 2.23 in the second equality. Applying Lemma 1.1 to each entry of $e^{-tA} e^{tA}$ shows that this matrix is equal to the identity, and the same proof shows this is also true for $e^{tA} e^{-tA}$.

We now prove the claim for general s and t. Fix $s \in \mathbb{R}$ and consider $e^{-(t+s)A} e^{tA} e^{sA}$ as a function of t. Since $e^{0 \cdot A}$ is the identity, this function is the identity when $t = 0$. Using the

product rule (Lemma 2.21), we can differentiate to obtain that

$$\frac{\partial}{\partial t}\left[e^{-(t+s)A}e^{tA}e^{sA}\right] = -Ae^{-(t+s)A}e^{tA}e^{sA} + e^{-(t+s)A}Ae^{tA}e^{sA}$$
$$= -Ae^{-(t+s)A}e^{tA}e^{sA} + Ae^{-(t+s)A}e^{tA}e^{sA} = 0$$

where we used Lemma 2.23 in the second equality. This implies the claim by similar reasoning as before. □

Theorem 2.27. *Let $A \in \mathcal{M}(n)$, let $I \subseteq \mathbb{R}$ be a non-trivial interval, and let $t_0 \in I$. Then $M : I \to \mathcal{M}(n)$ solves the ODE $M' = AM$ if and only if $M(t) = e^{(t-t_0)A}M(t_0)$ for every $t \in I$.*

Proof of Theorem 2.27. The fact that $e^{(t-t_0)A}M(t_0)$ solves the ODE for each $M(t_0) \in \mathcal{M}(n)$ follows from Lemmas 2.21 and 2.24. Now suppose that $M : I \to \mathcal{M}(n)$ is an arbitrary solution. We can use the matrix product rule Lemma 2.21 to calculate

$$(e^{-tA}M(t))' = -Ae^{-tA}M(t) + e^{-tA}AM(t) = -Ae^{-tA}M(t) + Ae^{-tA}M(t) = 0,$$

where we used Lemma 2.23 in the second equality. Applying Lemma 1.1 to each entry of $e^{-tA}M(t)$ implies that $e^{-tA}M(t) = e^{-t_0A}M(t_0)$ for every $t \in I$, and the claim follows by multiplying both sides on the left by e^{tA}. □

Remark 2.28. It follows that if $M(t_0)$ commutes with A then M also solves the ODE $M' = MA$. This is always the case in our application to phase space representations of constant coefficient linear ODEs, where $M(t_0)$ is the identity. In general, the solutions to the ODE $M' = MA$ are of the form $M(t_0)e^{(t-t_0)A}$.

Corollary 2.29. *Consider the constant-coefficient, homogeneous linear ODE*

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

which can be written in matrix notation as $\mathbf{f}' = A\mathbf{f}$. For each non-trivial interval $I \subseteq \mathbb{R}$ and $t_0 \in I$, the function $f : I \to \mathbb{R}$ solves the ODE if and only if

$$\mathbf{f}(t) = e^{(t-t_0)A}\mathbf{f}(t_0)$$

for every $t \in I$, where $=(f^{(n-1)}, f^{(n-2)}, \ldots, f)$.

Proof. As usual, we first need to check that functions of the given form solve the ODE. If $\mathbf{x}_0 \in \mathbb{R}^n$, we have by the product rule that

$$\frac{d}{dt}\left(e^{(t-t_0)A}\mathbf{x}_0\right) = \left(\frac{d}{dt}e^{(t-t_0)A}\right)\mathbf{x}_0 + e^{(t-t_0)A}\frac{d}{dt}\mathbf{x}_0 = Ae^{(t-t_0)A}\mathbf{x}_0,$$

so that $\mathbf{f} = e^{(t-t_0)A}\mathbf{x}_0$ is indeed a solution to the ODE. To prove that these are the only solutions, we can take an arbitrary phase-space solution $\mathbf{f}$ and differentiate $e^{-tA}\mathbf{f}$; the details are similar to before. □

The holomorphic functional calculus and the heat equation

The fact that we can sensibly apply functions defined through power series to matrices and preserve many of the properties that make these functions interesting as functions $\mathbb{R} \to \mathbb{R}$ (as we have just done with exponentials) has many applications. One further important example is the power series $(1-x)^{-1} = \sum_{n=0}^{\infty} x^n$, which is explored in exercise 17. One can also do similar things with "linear operators" on infinite dimensional spaces, like the differentiation operator $f \mapsto f'$ or the Laplacian $f \mapsto \Delta f$. This is known as the **holomorphic functional calculus**.

One paradigmatic example where these ideas arise is in the **heat equation** $\frac{\partial}{\partial t}u(t,x) = \Delta u(t,x)$. This is a PDE describing the time evolution of the function $u(t,\cdot)$, which describes the distribution of heat in a material. Appropriate infinite-dimensional versions of the ideas to those we have just explored let us express the solutions of the heat equation as $u(t,\cdot) = e^{t\Delta}u(0,\cdot)$, where the exponential $e^{t\Delta}$ is known as the **heat kernel** and can be expressed in terms of the density of the *Gaussian distribution* (as you may be familiar with from probability and statistics). Exponentials of linear maps on function spaces also arise in solutions to the *wave equation* and the *Schrodinger equation*, making the holomorphic functional calculus very important in mathematically rigorous treatments of quantum mechanics.

Exercise 17. Prove that if $A \in \mathcal{M}(n)$ with $\|A\|_{\text{op}} < 1$ then $I - A$ is invertible with inverse $\sum_{k=0}^{\infty} A^k$.

Exercise 18. Let $A : [0,\infty) \to \mathcal{M}(n)$ be a differentiable function. Prove that A satisfies the semigroup property stating that $A(0)$ is the identity and $A(t+s) = A(t)A(s)$ for all $s,t \geq 0$ if and only if there exists $L \in \mathcal{M}(n)$ such that $A(t) = e^{tL}$ for every $t \geq 0$.

Semigroups and their generators

There are versions of the fact proven in exercise 18 that do not assume differentiability of the semigroup and that apply in various infinite-dimensional settings: Look up "Hille–Yosida theorem" and "Stone's theorem on one-parameter unitary groups" on Wikipedia if you are interested. These theorems are important in mathematical quantum mechanics.

Exercise 19. Let $n \geq 1$. Given an $n \times n$ matrix $A \in \mathcal{M}(n)$, define

$$\sin(A) = \sum_{n=0}^{\infty} (-1)^n \frac{A^{2n+1}}{(2n+1)!} \qquad \text{and} \qquad \cos(A) = \sum_{n=0}^{\infty} (-1)^n \frac{A^{2n}}{(2n)!}.$$

1. Prove that these infinite series are well-defined, that $\sin(tA)$ and $\cos(tA)$ define differentiable functions of t for each $A \in \mathcal{M}(n)$, and that these functions satisfy the identities

$$\frac{d}{dt}\sin(tA) = A\cos(tA) \qquad \text{and} \qquad \frac{d}{dt}\cos(tA) = -A\sin(tA)$$

for every $A \in \mathcal{M}(n)$.

2. (Matrix Pythagoras.) Prove that $\sin(A)^2 + \cos(A)^2$ is equal to the identity matrix for every matrix A.

3. Prove that every solution to the second-order matrix ODE $M'' = -AM$ is of the form $\cos((t-t_0)A)M(t_0) + \sin((t-t_0)A)M'(t_0)$.

Doing computations with matrix exponentials. We have now expressed the solutions to constant-coefficient homogeneous linear ODEs in terms of matrix exponentials. Arguably, however, this is only useful insofar as we can actually compute matrix exponentials!

The first case to consider is when the matrix A is **diagonal**, meaning that all its non-diagonal entries are zero. In this case, A^k is also diagonal for every $k \geq 1$ with entries $A^k_{i,i} = (A_{i,i})^k$, so that the matrix exponential e^{tA} is also diagonal with entries $(e^{tA})_{i,i} = e^{tA_{i,i}}$. Unfortunately the matrices arising in our original linear ODE problem are never diagonal! This is not really a problem, however, provided that they are at least **diagonalizable**. Recall that this means that there exists an invertible matrix C and a diagonal matrix D (i.e., a matrix all of whose non-diagonal entries are zero) such that $A = CDC^{-1}$. Indeed, if we can find such matrices C and D then we have by induction that $A^k = CDC^{-1}CD^{k-1}C^{-1} = CD^kC^{-1}$ for every $k \geq 1$ and hence that

$$e^{tA} = \sum_{k=0}^{\infty} \frac{t^k A^k}{k!} = \sum_{k=0}^{\infty} \frac{t^k CD^kC^{-1}}{k!} = C\left[\sum_{k=0}^{\infty} \frac{t^k D^k}{k!}\right] C^{-1} = Ce^{tD}C^{-1},$$

where we used that the map $M \mapsto CMC^{-1}$ is continuous on $\mathcal{M}(n)$ to pull the conjugation outside of the infinite series in the second equality. (Note that this equality holds for any two conjugate matrices; we haven't yet used that D is diagonal.) Since e^{tD} is easy to compute, e^{tA} is reasonably easy also.

But when are matrices diagonalizable? As you will recall from Math 1, a matrix $A \in \mathcal{M}(n)$ is diagonalizable if and only if there exists a basis of $\mathbb{C}^n$ consisting of **eigenvectors**[10] for A, i.e., vectors x such that $Ax = \lambda x$ for some $\lambda \in \mathbb{C}$, where λ is known as the **eigenvalue** associated to the eigenvector x. Indeed, we can think of the diagonal matrix D as expressing the same linear transformation as A in terms of the basis of eigenvectors rather than the basis we started with: If $\{v_i\}$ is a basis of eigenvectors with eigenvalues $\{\lambda_i\}$ and C is the

[10] "Eigen" is a German word that is etymologically related to "own" in English. In this context it means something more like "characteristic".

matrix describing the linear map that sends the standard basis vector e_i (that has 1 in its ith coordinate and 0 everywhere else) to v_i then we can write $A = CDC^{-1}$ with $D_{i,i} = \lambda_i$.

Exercise 20. Let $a_0, \dots, a_{n-1}$ be constants and consider the matrix A defined by $A_{1,j} = -a_{n-j}$ and $A_{i,j} = \mathbb{1}(j = i-1)$ for each $i \neq 1$, as arises in the phase-space representation of a constant coefficient linear ODE. Prove that the eigenvalues of A coincide with the roots of the polynomial $\lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_1\lambda + a_0 = 0$.

Exercise 21. Let A be a matrix that encodes a constant coefficient linear ODE of the form $f^{(n)} + a_{n-1}f^{(n-1)} + \cdots + a_1 f' + a_0 f = 0$ as $\mathbf{f}' = A\mathbf{f}$. Prove that if $x \in \mathbb{C}^n$ is an eigenvector of A with eigenvalue λ then $x = (\lambda^{n-1}c, \lambda^{n-2}c, \dots, \lambda c, c)$ for some $c \in \mathbb{C}$. Deduce that A is diagonalizable if and only if it has n distinct eigenvalues.

Let's see how we can use this theory to compute the solution to an ODE in a simple example.

Example 2.30. The second-order ODE $f'' = -f$ can be written as a first-order ODE in phase space as

$$\begin{pmatrix} f' \\ f \end{pmatrix}' = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} f' \\ f \end{pmatrix}.$$

We can compute that this matrix has eigenvalues $+i$ and $-i$ with eigenvectors $(i, 1)$ and $(1, i)$ respectively, and can therefore be diagonalized

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} i & 1 \\ 1 & i \end{pmatrix} \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} \begin{pmatrix} i & 1 \\ 1 & i \end{pmatrix}^{-1} = \frac{1}{2} \begin{pmatrix} i & 1 \\ 1 & i \end{pmatrix} \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} \begin{pmatrix} i & -1 \\ -1 & i \end{pmatrix}.$$

As such, we have that

$$\begin{aligned} \exp\left[t \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \right] &= -\frac{1}{2} \begin{pmatrix} i & 1 \\ 1 & i \end{pmatrix} \begin{pmatrix} e^{it} & 0 \\ 0 & -e^{-it} \end{pmatrix} \begin{pmatrix} i & -1 \\ -1 & i \end{pmatrix} \\ &= \begin{pmatrix} \frac{e^{it}+e^{-it}}{2} & \frac{e^{it}-e^{-it}}{2i} \\ -\frac{e^{it}-e^{-it}}{2i} & \frac{e^{it}+e^{-it}}{2} \end{pmatrix} = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix}. \end{aligned}$$

Note that since the linear transformations

$$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \qquad \text{and } x + iy \mapsto i(x + iy)$$

are *the same map* after identifying $\mathbb{R}^2$ and $\mathbb{C}$ as real vector spaces by the isomorphism $(x, y) \mapsto x + iy$, the expression

$$\exp\left[t \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \right] = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix}$$

is really just Euler's formula

$$e^{it} = \cos(t) + i\sin(t)$$

in disguise! It follows that every solution to our ODE is of the form

$$\begin{pmatrix} f'(t) \\ f(t) \end{pmatrix} = \begin{pmatrix} \cos(t-t_0) & \sin(t-t_0) \\ -\sin(t-t_0) & \cos(t-t_0) \end{pmatrix} \begin{pmatrix} f'(t_0) \\ f(t_0) \end{pmatrix}.$$

This example already illustrates a general principle: When solving real-valued ODEs, we may have to diagonalize with complex eigenvalues, but these eigenvalues must always come in complex conjugate pairs, and ultimately lead to trig functions appearing in our ODE solutions as exactly as they did here.

Remark 2.31. As we saw in this example, we don't actually have to compute the matrix C in the diagonalization $A = CDC^{-1}$ if we only want to find *the space of all solutions*. When A is diagonalizable, it will always have a basis of solutions of the form

$$\begin{aligned} \{e^{\lambda t} : \lambda \text{ a real eigenvalue}\} \cup \{e^{\alpha t}\cos(\beta t) : \alpha \pm i\beta \text{ a complex conjugate pair of eigenvalues}\} \\ \cup \{e^{\alpha t}\sin(\beta t) : \alpha \pm i\beta \text{ a complex conjugate pair of eigenvalues}\}, \end{aligned}$$

so we know what the space of solutions is as soon as we've found the eigenvalues. On the other hand, the matrix C is telling us how to pick out a particular element of this space as a function of the initial conditions when solving the relevant IVP, since

$$\mathbf{f}(t) = e^{(t-t_0)A}\mathbf{f}(t_0) = Ce^{(t-t_0)D}C^{-1}\mathbf{f}(t_0).$$

Unfortunately, not every matrix is diagonalizable. Moreover, while a 'generic' matrix *is* diagonalizable, non-diagonalizable matrices turn out to be very important in ODE. Indeed, if we consider the simplest kth order ODE of all, namely

$$f^{(k)} = 0$$

then the associated matrix

$$A = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 & 0 \end{pmatrix}$$

is not diagonalizable! Indeed, this matrix is **nilpotent**, meaning that there exists $m \geq 1$ such that $A^m = 0$ (in this case $m = k+1$). While nilpotent matrices are not diagonalizable (unless they are the zero matrix), they are still easy to compute the exponential of: If N is nilpotent

with $N^m = 0$ then

$$e^{tN} = \sum_{i=0}^{\infty} \frac{t^i}{i!} N^i = \sum_{i=0}^{m-1} \frac{t^i}{i!} N^i$$

so that the matrix exponential e^{tN} is really just a polynomial! As such, if we exponentiate the matrix A encoding the linear ODE $f^{(k)} = 0$ we get that A^m is the matrix with ones on the mth subdiagonal and zeros everywhere else, so that

$$e^{tA} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ t & 1 & 0 & \cdots & 0 & 0 & 0 \\ \frac{1}{2}t^2 & t & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{(k-1)!}t^{k-1} & \frac{1}{(k-2)!}t^{k-2} & \frac{1}{(k-3)!}t^{k-3} & \cdots & t & 1 & 0 \\ \frac{1}{k!}t^k & \frac{1}{(k-1)!}t^{k-1} & \frac{1}{(k-2)!}t^{k-2} & \cdots & \frac{1}{2}t^2 & t & 1 \end{pmatrix}.$$

As expected, the solution $\mathbf{f}(t) = e^{(t-t_0)A}\mathbf{f}(t_0)$ agrees with our usual way of solving this ODE.

Now, it turns out that every matrix can be written as the *sum* of a diagonalizable matrix and a nilpotent matrix, letting us compute the exponential of any matrix in a reasonable way. Indeed, every matrix A can be written in the **Jordan normal form** $A = CJC^{-1}$ where C is invertible and J is of the form

$$J = \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_r \end{pmatrix} \qquad \text{where } J_i = \begin{pmatrix} \lambda_i & 1 & & & \\ & \lambda_i & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda_i & 1 \\ & & & & \lambda_i \end{pmatrix},$$

for some $\lambda_1, \ldots, \lambda_r \in \mathbb{C}$, where all the parts of the matrices left blank represent entries that are all zero. In particular, J can be written as the sum $J = D + N$ where D is diagonal, N is nilpotent, and D and N commute with each other. Once we have expressed a matrix $A = CJC^{-1}$ in Jordan normal form we can write

$$e^{tA} = Ce^{tJ}C^{-1} = C \begin{pmatrix} e^{tJ_1} & & \\ & \ddots & \\ & & e^{tJ_r} \end{pmatrix} C^{-1}.$$

To compute the exponential of each Jordan block e^{tJ_i} we will use the following lemma, whose proof is left as an exercise.

Exercise 22. If A and B commute then $e^{tA}e^{tB} = e^{t(A+B)}$ for every $t \in \mathbb{R}$.

If the Jordan block J_i has length k_i it can be written as the sum of $\lambda_i I_{k_i}$, where I_k is the

identity matrix of side-length k, and the standard nilpotent matrix of side-length k_i,

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

which satisfies

$$e^{tN_i} = \begin{pmatrix} 1 & t & \frac{1}{2}t^2 & \cdots & \frac{1}{(k-2)!}t^{k-2} & \frac{1}{(k-1)!}t^{k-1} & \frac{1}{k!}t^k \\ 0 & 1 & t & \cdots & \frac{1}{(k-3)!}t^{k-3} & \frac{1}{(k-2)!}t^{k-2} & \frac{1}{(k-1)!}t^{k-1} \\ 0 & 0 & 1 & \cdots & \frac{1}{(k-4)!}t^{k-4} & \frac{1}{(k-3)!}t^{k-3} & \frac{1}{(k-2)!}t^{k-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 & t \\ 0 & 0 & 0 & \cdots & 0 & 0 & 1 \end{pmatrix}.$$

Since N_{k_i} and $\lambda_i I_{k_i}$ commute, we have that

$$e^{tJ_i} = e^{t\lambda_i I_{k_i}} e^{tN_{k_i}} = e^{t\lambda_i} \begin{pmatrix} 1 & t & \frac{1}{2}t^2 & \cdots & \frac{1}{(k-2)!}t^{k-2} & \frac{1}{(k-1)!}t^{k-1} & \frac{1}{k!}t^k \\ 0 & 1 & t & \cdots & \frac{1}{(k-3)!}t^{k-3} & \frac{1}{(k-2)!}t^{k-2} & \frac{1}{(k-1)!}t^{k-1} \\ 0 & 0 & 1 & \cdots & \frac{1}{(k-4)!}t^{k-4} & \frac{1}{(k-3)!}t^{k-3} & \frac{1}{(k-2)!}t^{k-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 & t \\ 0 & 0 & 0 & \cdots & 0 & 0 & 1 \end{pmatrix},$$

and we can express e^{tA} in terms of blocks of this form conjugated by the matrix C.

Exercise 23. Continuing from Exercise 21, prove that if we write A in Jordan normal form then each Jordan block has a different associated eigenvalue. Prove moreover that the size of each Jordan block is equal to the multiplicity of the corresponding root of the polynomial $\lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_0 = 0$. (That is, if λ_0 is an eigenvalue of A, then the size of the Jordan block with eigenvalue λ is equal to the largest n such that $(\lambda - \lambda_0)^n$ divides the polynomial $\lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_0 = 0$.)

Note that if λ is an eigenvalue of the matrix A with eigenvector x then $\mathbf{f} = e^{t\lambda}x$ must be a solution to the ODE $\mathbf{f}' = A\mathbf{f}$. Writing this back in our original ODE notation and using that $(e^{t\lambda})^{(m)} = \lambda^m e^{t\lambda}$ for each $m \geq 0$, this means that λ is a solution to the polynomial

$$\lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_1\lambda + a_0 = 0.$$

Conversely, if λ solves this equation then $e^{\lambda t}$ is a solution to the ODE and λ is an eigenvalue

of A. When this polynomial has fewer than n distinct roots, A must be non-diagonalizable by the above exercise, and $\mathbf{f}' = A\mathbf{f}$ will therefore admit solutions that are either polynomials or products of polynomials and (complex) exponentials.

Example 2.32. Consider the second-order homogeneous linear ODE

$$f'' + 2\zeta f' + f = 0.$$

In matrix notation this ODE reads

$$\mathbf{f}' = \begin{pmatrix} -2\zeta & -1 \\ 1 & 0 \end{pmatrix} \mathbf{f}.$$

We can compute the eigenvalues of this matrix by solving

$$\det \begin{pmatrix} -2\zeta - \lambda & -1 \\ 1 & -\lambda \end{pmatrix} = \lambda^2 + 2\zeta\lambda + 1 = 0,$$

so that

$$\lambda = -\zeta \pm \sqrt{\zeta^2 - 1}.$$

If $|\zeta| > 1$ then we have two distinct real eigenvalues, and so there must exist an invertible matrix C such that

$$\exp\left[t \begin{pmatrix} -2\zeta & -1 \\ 1 & 0 \end{pmatrix}\right] = C \begin{pmatrix} e^{-(\zeta - \sqrt{\zeta^2-1})t} & 0 \\ 0 & e^{-(\zeta + \sqrt{\zeta^2-1})t} \end{pmatrix} C^{-1}.$$

Without needing to actually compute what these matrices C are, we deduce that every solution is of the form

$$f = \alpha e^{-(\zeta - \sqrt{\zeta^2-1})t} + \beta e^{-(\zeta + \sqrt{\zeta^2-1})t}$$

for some complex numbers α and β such that the right hand side is real for every $t \in \mathbb{R}$, which is the case if and only if α and β are both real.

Similarly, if $|\zeta| < 1$ then we have two distinct complex eigenvalues $\lambda = -\zeta \pm i\sqrt{1 - \zeta^2}$ coming in a conjugate pair, and again we have that every solution is of the form

$$f = \left(\alpha e^{it\sqrt{1-\zeta^2}} + \beta e^{-it\sqrt{1-\zeta^2}}\right) e^{-\zeta t}$$

for some complex numbers α and β such that the right hand side is real for every $t \in \mathbb{R}$, which now is the case if and only if α and β are complex conjugates of each other, so that $\alpha = x + iy$ and $\beta = x - iy$ for some $x, y \in \mathbb{R}$. Recognizing this sum of complex-conjugate exponentials as a sum of consine and sine functions, it follows that every solution is of the

form

$$f = \left[\tilde{\alpha}\cos(t\sqrt{1-\zeta^2}) + \tilde{\beta}\sin(t\sqrt{1-\zeta^2})\right] e^{-\zeta t}$$

for some real numbers $\tilde{\alpha}, \tilde{\beta}$.

Finally, suppose that $|\zeta| = 1$, so that $\lambda = -\zeta$ is the only eigenvalue. In this case there must exist an invertible matrix C such that

$$\begin{pmatrix} -2\zeta & -1 \\ 1 & 0 \end{pmatrix} = C \begin{pmatrix} -\zeta & 1 \\ 0 & -\zeta \end{pmatrix} C^{-1},$$

so that

$$\exp\left[t\begin{pmatrix} -2\zeta & -1 \\ 1 & 0 \end{pmatrix}\right] = e^{-\zeta t} C \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} C^{-1}.$$

Thus, again without having to actually compute C, we can deduce that every solution must be of the form

$$f = (\alpha + \beta t)e^{-\zeta t}$$

for some complex numbers α and β making the right hand side real for every $t \in \mathbb{R}$, which is the case if and only if α and β are both real.

A spring paradox? Before moving on, let us point out a hidden subtlety with what we have just done. When $\zeta \geq 0$, the equation $f'' + 2\zeta f' + f = 0$ is the equation of motion of a damped spring of mass and spring constant 1, with ζ describing the strength of the damping (which you can think of as describing how 'rigid' the spring is). We have seen that our solution to this equation looks very different depending on whether ζ is smaller than 1, equal to 1, or larger than 1. Suppose we release our spring at time zero with $f(0) > 0$ and $f' = 0$. When $0 < \zeta < 1$, the spring oscilates around its resting state of zero, with oscillations of decaying amplitude, while when $\zeta \geq 1$ it decays directly to zero without ever attaining a negative displacement; the $\zeta > 1$ and $\zeta = 1$ cases are qualitatively similar but with a quantitatively different form of decay at the special value $\zeta = 1$.

Thus, our algebraic approach to solving the equations seems to suggest that these three regimes are all very different in some sense. On the other hand, it seems likely that if we performed this experiment with springs that had $\zeta = 0.99999999$, $\zeta = 1$, and $\zeta = 1.00000001$ we would have a hard time telling the difference between the three springs without specialist equipment. Of course our analysis above had hard math to back it up, whereas the thought experiment we are now entertaining is just an appeal to intuition, but it is worth thinking about what is going on here, and whether the two things are really in tension with each other or not.

The resolution to this 'paradox' – that the solutions to the ODE should depend continuously on the parameter ζ but it seems that the solutions we have given do not – comes from the fact that the matrix C^{-1}, *which we did not compute*, **does not depend continuously**

on ζ! Indeed, it can't possibly depend continuously on ζ since if it did we would get that

$$C^{-1}\begin{pmatrix} -\zeta & 1 \\ 0 & -\zeta \end{pmatrix} C = \begin{cases} \begin{pmatrix} -(\zeta - \sqrt{\zeta^2-1}) & 0 \\ 0 & -(\zeta + \sqrt{\zeta^2-1}) \end{pmatrix} & \zeta \neq 1 \\ \begin{pmatrix} -\zeta & 1 \\ 0 & -\zeta \end{pmatrix} & \zeta = 1 \end{cases}$$

is continuous in ζ, which it isn't. Because we used the matrix C^{-1} to describe our solutions, it is only natural that *our description* has an abrupt change as ζ passes through the special value of 1, but this is not inconsistent with an observer making imperfect measurements of the system being unable to tell the difference between a value of ζ very slightly smaller than 1 or very slightly larger than 1.

Exercise 24. For each $\zeta \geq 0$, let f_ζ be the solution to the ODE $f'' + 2\zeta f' + f = 0$ with $f_\zeta(0) = 1$ and $f'_\zeta(0) = 0$. Compute f_ζ and prove that the function $\zeta \mapsto f_\zeta|_{[0,1]}$ defines a continuous function from $[0, \infty)$ to $C([0,1], \mathbb{R})$.

Unfortunately, the solution to the matrix ODE $M' = AM$ in terms of matrix exponentials does not generalize readily to the setting of *non-constant* A. There is one nice condition under which it does work:

Exercise 25. Let $I \subseteq \mathbb{R}$ be an open non-trivial interval, let $t_0 \in I$ and let $A : I \to \mathcal{M}(n)$ be continuous. Prove that if $A(t)$ commutes with $\int_{t_0}^t A(s)\,\mathrm{d}s$ for every $s \in I$ then every solution the matrix ODE $M'(t) = A(t)M(t)$ is of the form

$$M(t) = e^{\int_{t_0}^t A(s)\,\mathrm{d}s} M(t_0)$$

for every $t \in I$.

Exercise 26 (Bonus problem)**.** In this question, we will investigate what happens when we slightly relax the assumption of commutativity from the previous exercise.

a. Given two martrices A and B, the **commutator** $[A, B]$ is defined by $[A, B] = AB - BA$. A linear subspace L of $\mathcal{M}(n)$ is said to be a **Lie algebra** if it is closed under commutators, meaning that if $A, B \in L$ then $[A, B] \in L$. Prove that the space $\mathcal{N}(n) \subset \mathcal{M}(n)$ of n by n matrices that are 0 on and below the diagonal is a Lie algebra.

b. We say that a Lie algebra $\mathcal{N} \subseteq \mathcal{M}(n)$ is **nilpotent of step at most** 1 if $[A, [B, C]] = 0$ for every $A, B, C \in \mathcal{N}$. (Equivalently, if every element of the algebra commutes with every commutator formed from elements of the algebra.) Prove that the Lie algebra $\mathcal{N}(3)$ of 3 by 3 matrices with zeros on and below the diagonal is nilpotent of step at most 1.

c. Let $A : \mathbb{R} \to \mathcal{N}$ be a differentiable function from $\mathbb{R}$ to some Lie algebra $\mathcal{N} \subseteq \mathcal{M}(n)$ that is nilpotent of step at most 1. Prove that

$$\frac{d}{dt} e^{A(t)} = \left(A' + \frac{1}{2}[A, A'] \right) e^{A(t)}.$$

(Hint: Using that $\mathcal{N}$ is nilpotent of step at most 1, find a simple expression for $A^n B A^m - B A^{n+m}$ for $A, B \in \mathcal{N}$ and integers $n, m \geq 0$.)

d. Let $A : \mathbb{R} \to \mathcal{N}$ be a continuous function from $\mathbb{R}$ to some Lie algebra $\mathcal{N} \subseteq \mathcal{M}(n)$ that is nilpotent of step at most 1. Prove that the unique solution to the matrix ODE $M'(t) = A(t)M(t)$ with $M(0)$ equal to the identity is given by

$$M(t) = \exp\left[\int_0^t A(s)\,\mathrm{d}s - \frac{1}{2} \int_0^t \left[\int_0^s A(u)\,\mathrm{d}u, A(s) \right] \mathrm{d}s \right].$$

(Analogous facts are also true for step-s nilpotent Lie algebras, where the iterated commutator $[A_1, [A_2, [A_3, \cdots A_s, [A_{s+1}, A_{s+2}]]] \cdots] = 0$ vanishes for all $A_1, \ldots, A_{s+2}$. Examples include the space of upper-triangular matrices in $d+2$ dimensions with zeros on the diagonal. For such algebras, solutions to the relevant ODEs also involve more iterated commutators. Without any nilpotency assumptions, one can still write the solution *for small times* as an infinite series of iterated commutators. Look up the Baker-Campbell-Hausdorff formula if you want to learn more.)

2.6 The cookbook solution

I will now describe the standard algorithm for computing the solution to a constant coefficient homogeneous linear ODE, which gives the same answer as going through matrix exponentiation but can easily be performed "robotically" without knowing where the method comes from. Suppose we want to solve a constant coefficient linear ODE

$$f^{(n)} + a_{n-1} f^{(n-1)} + \cdots + a_0 f \equiv 0.$$

We can do this via the following procedure.

1. First look for solutions of the form $f = e^{\lambda t}$, where λ is a complex number. This function will solve the ODE if and only if λ is a root of the polynomial

$$\lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_1 \lambda + a_0 = 0.$$

(In fact the roots of this polynomial are precisely the eigenvalues of the matrix A in the phase space representation $\mathbf{f}' = A\mathbf{f}$.) If this polynomial has complex roots, they must come in complex-conjugate pairs $\lambda = \alpha \pm i\beta$, and the corresponding real solutions are $e^{\alpha t}\cos(\beta t)$ and $e^{\alpha t}\sin(\beta t)$.

2. If we found n distinct roots in the first step, then the relevant exponentials and trig functions span our whole space of real solutions, so we are done. If we do not have n distinct eigenvalues, the roots of the polynomial $\lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_1\lambda + a_0 = 0$ that have multiplicity more than 1 will lead to solutions that are multiples of polynomials and the (complex) exponentials written above, where the maximum degree of the polynomial is equal to one minus the multiplicity of the root.

Example 2.33. Let's consider the ODE $f^{(3)} = f$. The eigenvalues of the relevant matrix are the solutions to $\lambda^3 = 1$, i.e., the third roots of unity 1, $e^{i\frac{2}{3}\pi}$, and $e^{-i\frac{2}{3}\pi}$. Writing $e^{i\frac{2}{3}\pi} = -\frac{1}{2} - \frac{\sqrt{3}}{2}i$, we deduce that the solutions to this ODE are of the form

$$Ae^t + Be^{-t/2}\cos(\frac{\sqrt{3}}{2}t) + Ce^{-t/2}\sin(\frac{\sqrt{3}}{2}t).$$

Notice how different initial conditions can lead to very different large-time behaviours according to whether $A = 0$.

Example 2.34. Let's consider the ODE $f^{(3)} - 6f^{(2)} + 12f' - 8 = 0$. The eigenvalues of the relevant matrix are the solutions to $\lambda^3 - 6\lambda^2 + 12\lambda - 8 = 0$, which we can recognise as $(\lambda - 2)^3 = 0$. Thus, there is only one eigenvalue, 2, with multiplicity 3. Thus, the solutions to the ODE are all of the form $(A + Bt + Ct^3)e^{2t}$.

Example 2.35. Let's consider the ODE $f^{(8)} - 6f^{(7)} + 27f^{(6)} - 68f^{(5)} + 135f^{(4)} - 150f^{(3)} + 125f^{(2)} = 0$, whose associated polynomial can be factored $\lambda^2(\lambda - 1 + 2i)^3(\lambda - 1 - 2i)^3$. (This is a contrived example, and I would not expect you to factor this polynomial by hand.) Thus, there is one zero eigenvalue of multiplicity two and a complex conjugate pair of eigenvalues each of multiplicity three. Every solution is of the form

$$A + Bt + (C + Dt)e^t\cos(2t) + (E + Ft)e^t\sin(2t).$$

Exercise 27. Find all solutions $f : \mathbb{R} \to \mathbb{R}$ to the ODE $f^{(2024)} = f$.

Exercise 28. Find all solutions $f : \mathbb{R} \to \mathbb{R}$ to the ODE $f^{(4)} + f^{(2)} + f = 0$.

3 Existence, Uniqueness, and Regularity

Our next goal is to state and prove the most basic and important existence and uniqueness theorem for first-order ODEs, the *Picard-Lindelöf Theorem* (a.k.a. Picard's existence theorem, a.k.a. the Cauchy–Lipschitz theorem).

Theorem 3.1 (Picard-Lindelöf – local, first-order version)**.** *Let $\Omega \subseteq \mathbb{R}^{1+d}$ be an open set and let $F : \Omega \to \mathbb{R}^d$ be a continuous function for which there exists a constant M such that*

$$\|F(t, x_1) - F(t, x_2)\| \leq M\|x_1 - x_2\| \tag{3.1}$$

for every $(t, x_1), (t, x_2) \in \Omega$. Then for each $(t_0, x_0) \in \Omega$ there exists $\varepsilon > 0$ such that if I is a non-trivial closed interval containing t_0 of length at most ε then the ODE $f' = F(t, f)$ has a unique solution on I with $f(t_0) = x_0$.

Note that this theorem is "local" in the sense that it only guarantees existence and uniqueness of solutions on a possibly very small interval around the starting time. We will return later to the issue of how big we can actually take the domain of our solutions to be.

Before we start working towards the proof of this theorem, some remarks are in order about its statement. A function F defined on a subset Ω of $\mathbb{R}^d$ is said to be **Lipschitz** (or **Lipschitz continuous**) if there exists a constant M such that $\|F(x) - F(y)\| \leq M\|x - y\|$ for every $x, y \in \Omega$; the hypothesis of the Picard-Lindelöf theorem can be expressed more succinctly as the condition that $F(t, x)$ is continuous in t and Lipschitz in x. Note that if $F : \Omega \to \mathbb{R}^d$ is differentiable and has bounded partial derivatives in all directions (equivalently, bounded total derivative in the operator norm) then it automatically satisfies this condition.

Remark 3.2. To see that the Lipschitz condition is needed for the *uniqueness* part of the theorem to be true, consider the ODE $f' = |f|^{2/3}$, which has two solutions with $f(0) = 0$ given by $f \equiv 0$ and $f(t) \equiv \frac{1}{27}t^3$. For the *existence* part of the theorem it suffices for F to be continuous, a fact known as the *Peano existence theorem*. We will not prove the Peano existence theorem in this course.

Let us now start working towards the proof of the theorem. We will work with the equivalent *integral equation*

$$f(t) = x_0 + \int_{t_0}^{t} F(s, f(s))\,\mathrm{d}s. \tag{3.2}$$

The basic idea of the proof will be to construct a solution to the equation via what is called **Picard iteration**: We start with $f_0 \equiv x_0$ and, at each step $k \geq 1$, construct a new function f_k satisfying

$$f_k'(t) = F(t, f_{k-1}(t)) \text{ and } f_k(t_0) = x_0$$

by setting

$$f_k(t) = x_0 + \int_{t_0}^{t} F(s, f_{k-1}(s))\,\mathrm{d}s.$$

It is plausible that if this sequence of functions converges to some function f then f should satisfy (3.2) as required, and we will see that this is indeed the case. Moreover, we will see that the *function on the space of functions* defined by

$$f \mapsto x_0 + \int_{t_0}^{t} F(s, f(s))\,\mathrm{d}s \tag{3.3}$$

tends to "push functions closer together", so that if we were to iterate this map starting at two different functions then they should converge to the *same* fixed point. But any solution to the ODE is a fixed point of the map, and the solution to the ODE must therefore be unique!

Filling out the details of the proof will require introducing a few basic concepts that are very important throughout analysis. If you take more courses in analysis you will develop many of these concepts (metric spaces, completeness, Banach spaces, ...) in a systematic way. Since that is not the purpose of this course, we will instead develop the relevant theory only in the specific context we need it, although in fact the general proofs are not really any different.

3.1 The space of continuous functions

In order to start implementing the proof we just sketched rigorously, the first step will be to introduce an appropriate space of functions on which to define the map (3.3), together with a suitable notion of what it means for a sequence of functions to converge in this space.

For this we will need a generalization of the extreme value theorem. A subset[11] V of $\mathbb{R}^n$ is said to be **closed** if whenever $(x_n)_{n\geq 0}$ is a sequence in V converging to some point $x \in \mathbb{R}^n$, then $x \in V$. (That is, a set is closed if it contains all its limit points.)

Exercise 29. Prove that a subset V of $\mathbb{R}^n$ is closed if and only if its complement $V \setminus \mathbb{R}^n$ is open. Prove that the only sets that are both open and closed in $\mathbb{R}^n$ are the whole set $\mathbb{R}^n$ and the empty set $\emptyset$.

A subset V of $\mathbb{R}^n$ is said to be **bounded** if there exists $C < \infty$ such that $V \subseteq \{x \in \mathbb{R}^n : \|x\| \leq C\}$. (Here we take $\|\cdot\|$ to be the Euclidean norm, but the choice does not affect the definition of being bounded.)

Theorem 3.3 (Multivariable extreme value theorem)**.** *Let V be a closed, bounded set in $\mathbb{R}^n$ and let $f : V \to \mathbb{R}$ be a continuous function. Then there exists $x_0 \in V$ such that $\sup\{f(x) : x \in V\} = f(x_0)$. In particular, $\sup\{f(x) : x \in V\}$ is finite.*

Proof. This proof is similar to that of the single-variable version and is left as an exercise. □

Exercise 30. Prove Theorem 3.3.

[11] It is traditional to denote open sets by U and closed sets by V. This does unfortunately clash with our traditional notation for a vector space, but there are only so many letters in the alphabet.

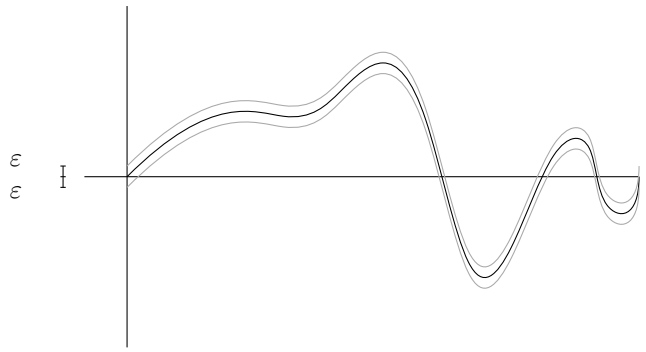


Figure 5: A continuous function f is graphed in black. A continuous function g satisfies $\|f-g\|_\infty \leq \varepsilon$ if and only if it stays between the two grey shifted copies of f. (Can you explain why the ribbon of allowed functions looks narrower when the slope of f is steeper?)

Remark 3.4. A more sophisticated way to state this theorem is that "closed, bounded subsets of $\mathbb{R}^n$ are compact". This is often known as the *Heine-Borel theorem.*

Let V be a closed, bounded set in $\mathbb{R}^n$ for some $n \geq 1$, and let $C(V, \mathbb{R}^m)$ be the space of continuous functions from V to $\mathbb{R}^m$. The **uniform norm** $\|\cdot\|_\infty$ on $C(V, \mathbb{R}^m)$ is the function $\|\cdot\|_\infty : C(V, \mathbb{R}^m) \to \mathbb{R}$ defined by

$$\|f\|_\infty = \sup_{x \in V} \|f(x)\|,$$

where $\|f(x)\|$ denotes the Euclidean norm of $f(x)$. The uniform norm $\|f\|_\infty$ is finite for every $f \in C(V, \mathbb{R}^m)$ by the extreme value theorem. Let's now check in detail that it really does define a norm on the vector space $C(V, \mathbb{R}^m)$. The only non-obvious thing we need to check is that the triangle inequality holds:

Lemma 3.5 (The triangle inequality)**.** *Given $n, m \geq 1$ and a closed, bounded set $V \subseteq \mathbb{R}^n$ the inequality*

$$\|f+g\|_\infty \leq \|f\|_\infty + \|g\|_\infty$$

holds for every $f, g \in C(V, \mathbb{R}^m)$.

Proof. We have that

$$\begin{aligned}\|f+g\|_\infty = \sup_{x \in V} \|f(x)+g(x)\| &\leq \sup_{x \in V} (\|f(x)\| + \|g(x)\|) \\ &\leq \sup_{x \in V} \|f(x)\| + \sup_{y \in \tilde{I}} \|g(y)\| = \|f\|_\infty + \|g\|_\infty,\end{aligned}$$

where the first inequality holds by the triangle inequality for the Euclidean distance on $\mathbb{R}^m$. □

We say that a sequence of functions $(f_n)_{n \geq 1}$ in $C(V, \mathbb{R}^m)$ **converges uniformly** to a function $f \in C(V, \mathbb{R}^m)$ if $\|f_n - f\|_\infty \to 0$ as $n \to \infty$. As before, we say that a sequence of

functions $(f_n)_{n\geq 1}$ in $C(V,\mathbb{R}^m)$ is a **Cauchy sequence** (with respect to $\|\cdot\|_\infty$) if for every $\varepsilon > 0$ there exists $N < \infty$ such that $\|f_n - f_m\|_\infty \leq \varepsilon$ for every $n, m \geq N$.

We will need to know that Cauchy sequences in $C(V,\mathbb{R}^m)$ always have limits.

Theorem 3.6. *Let $n, m \geq 1$ and let $V \subseteq \mathbb{R}^n$ be a closed, bounded set. If $(f_n)_{n\geq 1}$ is a Cauchy sequence in $C(V,\mathbb{R}^m)$ then there exists $f \in C(V,\mathbb{R}^m)$ such that f_n converges uniformly to f.*

Proof. First note that for each $x \in V$, $(f_n(x))_{n\geq 1}$ is a Cauchy sequence in $\mathbb{R}^m$ (because $\|f_n(x) - f_m(x)\| \leq \|f_n - f_m\|_\infty$ for every $x \in V$) and therefore converges to some limit, which we may denote $f(x)$. It suffices to prove that this defines a *continuous* function $f : V \to \mathbb{R}^m$ and that $\|f_n - f\| \to 0$ as $n \to \infty$. For each $x \in V$ we have that

$$\|f(x) - f_n(x)\| = \lim_{m\to\infty} \|f_m(x) - f_n(x)\|$$

and hence that for each $\varepsilon > 0$ there exists $N < \infty$ such that

$$\|f(x) - f_n(x)\| \leq \varepsilon \qquad \text{for every } n \geq N \text{ and } x \in V. \tag{3.4}$$

Let us now see why this implies continuity of f. Fix $x \in \mathbb{R}$ and $\varepsilon > 0$, and let n be such that $\|f(y) - f_n(y)\| \leq \varepsilon/3$ for every $y \in V$. Since f_n is continuous, there exists δ such that if $\|y - x\| \leq \delta$ then $\|f_n(y) - f_n(x)\| \leq \varepsilon/3$. Thus, if $y \in V$ is any point satisfying $\|y - x\| \leq \delta$ then

$$\|f(y) - f(x)\| \leq \|f(y) - f_n(y)\| + \|f_n(y) - f_n(x)\| + \|f_n(x) - f(x)\| \leq \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon.$$

Since $x \in V$ and $\varepsilon > 0$ were arbitrary this implies that f is continuous as claimed. Now that we know f is continuous, we can rewrite (3.4) equivalently as the statement that for each $\varepsilon > 0$ there exists $N < \infty$ such that $\|f - f_n\|_\infty \leq \varepsilon$ for every $n \geq N$, so that $\|f - f_n\|_\infty \to 0$ as $n \to \infty$. □

Remark 3.7. This theorem is usually stated as "spaces of continuous functions with the uniform norm are Banach spaces", where a *Banach space* is normed vector space that is *complete*, meaning that all its Cauchy sequences have limits. We avoid using this terminology here since we don't want to get into the general theory of such spaces. (You will do this if you take more courses in analysis.)

Exercise 31. A sequence of functions $(f_n)_{n\geq 1}$ from $[0,1]$ to $\mathbb{R}$ is said to converge **pointwise** to a function $f : [0,1] \to \mathbb{R}$ if $f_n(x) \to f(x)$ for every $x \in \mathbb{R}$. Prove that a pointwise limit of continuous functions need not be continuous.

Now that we have a notion of distance on the space of continuous functions, we can make sense of what it means for a function on this space to be continuous. If $V \subseteq \mathbb{R}^n$ is a closed, bounded set and A is a subset of $C(V,\mathbb{R}^m)$, we say that a function $\phi : A \to C(V,\mathbb{R}^m)$ is $\|\cdot\|_\infty$**-continuous** (or **continuous with respect to** $\|\cdot\|_\infty$) at a function $f \in A$ if for every

$\varepsilon > 0$ there exists $\delta > 0$ such that if $g \in A$ satisfies $\|g - f\|_\infty \leq \delta$ then $\|\phi(g) - \phi(f)\|_\infty \leq \varepsilon$, and say that ϕ is $\|\cdot\|_\infty$-continuous if it is continuous at every $f \in A$.

Lemma 3.8. *Let I be a non-trivial closed bounded interval in $\mathbb{R}$ and let t_0 be a point in I. For each $d \geq 1$ integration defines a function*

$$\begin{aligned} \textstyle\int : C(I, \mathbb{R}^d) &\longrightarrow C(I, \mathbb{R}^d) \\ f(t) &\longmapsto \displaystyle\int_{t_0}^{t} f(s)\,\mathrm{d}s \end{aligned}$$

that is continuous with respect to $\|\cdot\|_\infty$.

This proof will use the fact that

$$\left\| \int_a^b f(s)\,\mathrm{d}s \right\| \leq \int_a^b \big\|f(s)\big\|\,\mathrm{d}s$$

for any continuous function $f : [a, b] \to \mathbb{R}^d$. (Recall that the integral of a multivariable function $f : [a, b] \to \mathbb{R}$ is defined by integrating each coordinate separately.) If you haven't seen this inequality before you should convince yourself that it is true (e.g. using the definition of the Riemann integral).

Proof of Lemma 3.8. For each function $f \in C(I, \mathbb{R}^d)$ let $\int f$ be the function defined by $\int f(t) = \int_{t_0}^t f(s)\,\mathrm{d}s$. For each $f, g \in C(I, \mathbb{R}^m)$ we have that

$$\begin{aligned} \|\textstyle\int f - \int g\|_\infty &= \sup_{t\in I} \left\| \displaystyle\int_{t_0}^t f(s)\,\mathrm{d}s - \int_{t_0}^t g(s)\,\mathrm{d}s \right\| \\ &\leq \sup_{t\in I} \left| \int_{t_0}^t \|f(s) - g(s)\|\,\mathrm{d}s \right| \leq \sup_{t\in I} |t - t_0| \|f - g\|_\infty \leq |I| \cdot \|f - g\|_\infty \end{aligned}$$

where $|I|$ is the length of I, which is finite since I is bounded. (We have to put the absolute values in after the first inequality to deal with the case $t < t_0$.) This implies $\|\cdot\|_\infty$-continuity of $\int$: For each $\varepsilon > 0$, if $f, g \in C(I, \mathbb{R}^d)$ satisfy $\|f - g\|_\infty \leq \delta = \varepsilon/|I|$ then $\|\int f - \int g\|_\infty \leq \varepsilon$. □

Example 3.9. Let $C^1([0, 1], \mathbb{R})$ be the subset of $C([0, 1], \mathbb{R})$ consisting of those functions that are continuously differentiable. Then differentiation does *not* define a $\|\cdot\|_\infty$-continuous function $C^1([0, 1], \mathbb{R}) \to C([0, 1], \mathbb{R})$: The functions $\frac{1}{n}\sin(nx)$ converge uniformly to the zero function as $n \to \infty$ but their derivatives $(\frac{1}{n}\sin(nx))' = \cos(nx)$ do not converge uniformly to zero.

Term-by-term differentiation. We now have everything we need to go back and prove Theorem 2.25, which we left unproven in our discussion of matrix exponentials. Strictly speaking this requires versions of everything we have just done for functions taking values in an arbitrary finite-dimensional vector space, but this follows from the $\mathbb{R}^d$ version.

Proof of Theorem 2.25. This is just $\|\cdot\|_\infty$-continuity of integration in disguise! The same proof we just also implies that integration defines a continuous function from $C(I,V)$ to itself when V is any finite-dimensional vector space. That is, if we fix a base point $t_0 \in I$ then

$$f \mapsto \int_{t_0}^{t} f(s)\,\mathrm{d}s$$

defines a continuous function $C(I,V) \to C(I,V)$.

Now, the condition $\sum_{i=1}^{\infty} \|f_i^{(n)}\|_\infty < \infty$ implies that the partial sums $(\sum_{i=1}^{N} f_i^{(n)})_{N\geq 1}$ form a Cauchy sequence. The same proof we did earlier in the case $V = \mathbb{R}^d$ shows that every Cauchy sequence in $C(I,V)$ converges to a limit in $C(I,V)$, and we can call the limit of these partial sums g. Fix $t_0 \in I$. Since (iterated) integrals are $\|\cdot\|_\infty$-continuous and $(\sum_{i=1}^{N} f_i^{(n)})_{N\geq 1}$ converges to g, we must have that $\sum_{i=1}^{N} f_i^{(m)}(t) - \sum_{i=1}^{N} f^{(m)}(t_0)$ converges (with respect to $\|\cdot\|_\infty$) to the $(n-m)$th iterated integral of g for each $0 \leq n \leq m$ as $N \to \infty$. Since we also have that the partial sums $\sum_{i=1}^{N} f^{(m)}(t_0)$ converge for each $0 \leq m \leq n$ by assumption, it follows that, for each $0 \leq m < n$, the partial sums $\sum_{i=1}^{N} f_i^{(m)}(t)$ converge (with respect to $\|\cdot\|_\infty$) to a function whose $(n-m)$th derivative is g as claimed. □

Exercise 32. Let $C^1([0,1])$ be the space of continuously differentiable functions from $[0,1]$ to $\mathbb{R}$. Given $f \in C^1([0,1])$, define

$$\|f\|_{C^1} = |f(0)| + \|f'\|_\infty.$$

Prove that this is a norm. Prove that if $(f_n)_{n\geq 1}$ is a Cauchy sequence in $C^1([0,1])$ in the sense that

$$\lim_{n\to\infty} \sup_{m\geq n} \|f_m - f_n\|_{C^1} = 0$$

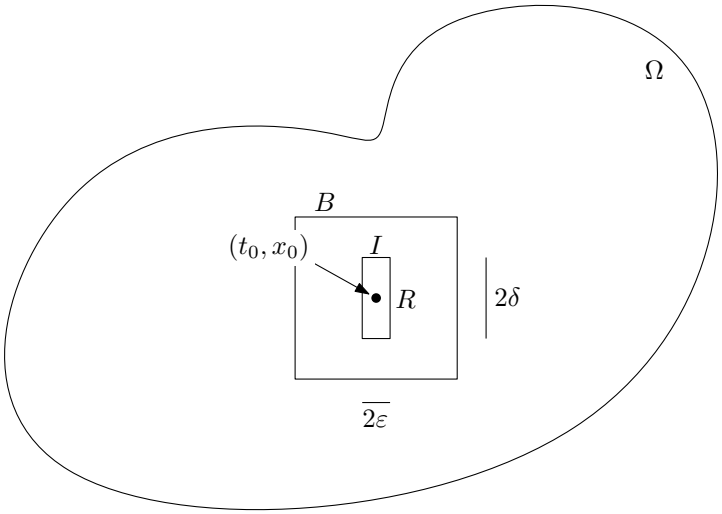
then there exists a function $f \in C^1([0,1])$ such that $\|f_n - f\|_{C^1} \to 0$.

3.2 Proof of Picard-Lindelöf

Proof of Picard-Lindelöf. Fix $(t_0, x_0) \in \Omega$. Since Ω is open, there exists $\delta > 0$ such that the box B of side-length 4δ centered at (t_0, x_0) is contained in Ω. Since this box is closed, there exists a constant C such that $\|F(t,x)\| \leq C$ for every $(t,x) \in B$ by Theorem 3.3. Let

$$\varepsilon = \min\left\{\delta, \frac{\delta}{2C}, \frac{1}{2M}\right\}.$$

The reason for this particular choice of ε will become apparent later in the proof; the basic point is that the integral transformation Γ that we will iterate to find our solution is 'more continuous' the smaller interval we take when defining it. Fix a closed interval I containing t_0 that has length at most ε; we need to prove that the IVP has a unique solution on I.



Remark 3.10. If you were coming up with this proof for the first time, you would keep ε as a variable and then figure out how small you need to take it at the end of the proof. When writing up the proof it's usually nicer to define it up front so that you are certain there's no circularity in the conditions defining how small it needs to be.

Step 1: Defining the map. Let R be the closed box of sidelength 2δ around x_0, so that $I \times R \subseteq B \subseteq \Omega$. Let

$$\mathscr{C} = \{f \in C(I, \mathbb{R}^d) : f(t_0) = x_0 \text{ and } f(t) \in R \text{ for every } t \in I\}.$$

We claim that if $f \in \mathscr{C}$ then the function Γf defined by

$$\Gamma f(t) = x_0 + \int_{t_0}^{t} F(s, f(s))\,\mathrm{d}s \qquad t \in I$$

also belongs to $\mathscr{C}$. Clearly Γf is continuous and satisfies $\Gamma f(t_0) = x_0$, so it suffices to prove that Γf takes values only in R. To see this, note that, since $\|F\| \leq C$ on B we have that

$$\|\Gamma f(t) - x_0\| \leq \int_{t_0}^{t} \|F(s, f(s))\|\,\mathrm{d}s \leq C|t - t_0| \leq \delta$$

for every $t \in I$ and hence that $\Gamma f(t) \in R$ for every $t \in I$ as claimed. (This is why we made sure $\varepsilon \leq \delta/C$ when choosing $\varepsilon > 0$.) Thus, we have a well-defined function

$$\begin{aligned} \Gamma : \mathscr{C} &\longrightarrow \mathscr{C} \\ f &\longmapsto x_0 + \int_{t_0}^{t} F(s, f(s))\,\mathrm{d}s. \end{aligned}$$

Our next goal is to prove that this map tends to "push functions closer together" in an appropriate sense.

Step 2: The contraction property. We claim that if $f_1, f_2 \in \mathscr{C}$ then

$$\|\Gamma f_1 - \Gamma f_2\|_\infty \leq \frac{1}{2}\|f_1 - f_2\|_\infty. \tag{3.5}$$

(The fact that we get $1/2$ here isn't very important; any constant strictly less than 1 would work.) Indeed, for each $t \in I$ we have by the Lipschitz assumption on F that

$$\begin{aligned}\|\Gamma f_1 - \Gamma f_2\| \leq \left| \int_{t_0}^t \|F(s, f_1(s)) - F(s, f_2(s))\| \, \mathrm{d}s \right| &\leq M \left| \int_{t_0}^t \|f_1(s) - f_2(s)\| \, \mathrm{d}s \right| \\ &\leq M(t - t_0)\|f_1 - f_2\|_\infty \leq \frac{1}{2}\|f_1 - f_2\|_\infty,\end{aligned}$$

where the final inequality holds since we took $\varepsilon \leq 1/2M$.

Step 3: A Cauchy sequence via Picard iteration. Let $f_0 \in \mathscr{C}$ be given by taking $f_0 \equiv x_0$. For each $n \geq 1$, we define $f_n \in \mathscr{C}$ by $f_n = \Gamma f_{n-1}$. We have by (3.5) that

$$\|f_{n+1} - f_n\|_\infty = \|\Gamma f_n - \Gamma f_{n-1}\|_\infty \leq \frac{1}{2}\|f_n - f_{n-1}\|_\infty$$

for each $n \geq 1$ and hence by induction that

$$\|f_{n+1} - f_n\|_\infty \leq 2^{-n}\|f_1 - f_0\|_\infty$$

for every $n \geq 0$. It follows by the triangle inequality that

$$\|f_m - f_n\| \leq \sum_{i=n}^{m-1} \|f_{i+1} - f_i\|_\infty \leq \sum_{i=n}^{m-1} 2^{-i}\|f_1 - f_0\|_\infty \leq 2^{-n+1}\|f_1 - f_0\|_\infty$$

and hence that $(f_n)_{n\geq 1}$ is a Cauchy sequence in $C(I, \mathbb{R}^m)$. Thus, there must exist some function $f \in C(I, \mathbb{R}^m)$ such that $\|f - f_n\|_\infty \to 0$ as $n \to \infty$, and this function f must also belong to $\mathscr{C}$ (proving this in detail is a nice simple exercise to get to grips with the $\|\cdot\|_\infty$ norm). Finally, we have that

$$\begin{aligned}\|\Gamma f - f\|_\infty &\leq \|\Gamma f - \Gamma f_n\|_\infty + \|\Gamma f_n - f_n\|_\infty - \|f_n - f\|_\infty \\ &\leq \|\Gamma f_n - f_n\|_\infty + \frac{3}{2}\|f - f_n\|_\infty\end{aligned}$$

for every $n \geq 1$, and since the right hand side tends to zero as $n \to \infty$ we must have that $\|\Gamma f - f\| = 0$ and hence that $\Gamma f = f$. In other words, f is a solution to the integral equation (3.2) and hence to the ODE $f' = F(t, f)$ on I with $f(t_0) = x_0$.

Step 4: Uniqueness. Suppose that $f_1, f_2 \in \mathscr{C}$ are both solutions to the ODE $f' = F(t, f)$ on $I = [t_0 - \varepsilon, t_0 + \varepsilon]$ with $f_1(t_0) = f_2(t_0) = x_0$. Then f_1 and f_2 both solve the integral equation (3.2) and hence are fixed points of Γ in the sense that $\Gamma f_1 = f_1$ and $\Gamma f_2 = f_2$. But

in this case we have that $\|f_1 - f_2\|_\infty = \|\Gamma f_1 - \Gamma f_2\|_\infty \leq \frac{1}{2}\|f_1 - f_2\|_\infty$ which holds if and only if $\|f_1 - f_2\|_\infty = 0$, if and only if $f_1 = f_2$. Thus, there is at most one solution $f \in \mathscr{C}$ to the ODE satisfying $f(t_0) = x_0$.

Unfortunately this is not *quite* what we wanted to prove: we also need to rule out the existence of solutions that *don't* belong to $\mathscr{C}$, i.e., that don't take values in R. Ruling out such a solution can be done similarly to how we proved that $\Gamma : \mathscr{C} \to \mathscr{C}$ took values in $\mathscr{C}$. Indeed, suppose that (I, f) is any solution to the integral equation (3.2). Since R is closed and f is continuous, we can consider the largest possible closed interval $[t_-, t_+] \subseteq I$ containing t_0 such that $f(t) \in R$ for every $t \in [t_-, t_+]$, defined by

$$t_- = \sup\{t \leq t_0 : t \notin I \text{ or } f(t) \notin R\}$$
$$t_+ = \inf\{t \geq t_0 : t \notin I \text{ or } f(t) \notin R\}.$$

Since $f(t) \in R$ for every $t \in [t_-, t_+]$, we have that if $t \in [t_-, t_+]$ then

$$\|f(t) - x_0\| = \left\| \int_{t_0}^{t} F(s, f(s))\,\mathrm{d}s \right\| \leq \int_{t_0}^{t} \big\|F(s, f(s))\big\|\,\mathrm{d}s \leq C|t - t_0| \leq \frac{\delta}{2}$$

since $\varepsilon \leq \delta/2C$. As such, neither $f(t_-)$ or $f(t_+)$ can belong to the boundary of R (which is centered at x_0 and has side lengths 2δ), and since f is continuous we must have that $t_- = \inf I$ and $t_0 = \sup I$, so that f takes values in R as desired. □

3.3 Higher-order Picard-Lindelöf

Recall Proposition 2.19, which explained how for each nth order ODE with d-dimensional solutions, there is a first-order ODE with nd-dimensional solutions whose solutions are naturally in bijection with those of the original ODE. This proposition has the following consequence when combined with Picard-Lindelöf.

Corollary 3.11 (Higher-order Picard-Lindelöf)**.** *Let $\Omega \subseteq \mathbb{R}^{1+nd}$ be a non-empty open set and let $F : I \times \Omega \to \mathbb{R}^d$ be a continuous function for which there exists a constant M such that*

$$\|F(t, \mathbf{x}) - F(t, \mathbf{y})\| \leq M\|\mathbf{x} - \mathbf{y}\| \tag{3.6}$$

for every $(t, \mathbf{x}), (t, \mathbf{y}) \in \Omega$. Then for each $(t_0, \mathbf{x}_0) \in \Omega$ there exists $\varepsilon > 0$ such that if I is a non-trivial closed interval of length at most ε containing t_0 then the ODE *$f^{(n)} = F(t, f)$ has a unique solution on I with $f^{(i)}(t_0) = x_{0,i}$ for each $0 \leq i \leq n - 1$.*

Checking that this corollary follows from the first-order version of Picard-Lindelöf is left as an exercise below.

Exercise 33. Let $I \subseteq \mathbb{R}$ be a non-trivial interval, let $n \geq 1$, and let $f : I \to \mathbb{R}$ and $g : I \to \mathbb{R}$ be n-times differentiable. Given a set $S \subseteq I$, we say that $t \in I$ is an **accumulation point** of S if there exists a sequence $(t_n)_{n\geq 1}$ in $S \setminus \{t\}$ such that $t_n \to t$ as $n \to \infty$.

1. Prove that if $S \subseteq I$ is such that $f(s) = g(s)$ for every $s \in S$ then $f^{(m)}(t) = g^{(m)}(t)$ for every accumulation point t of S in I and every $m \leq n$. (Warning: an accumulation point of S need not be an accumulation point of the set of accumulation points of S!)

2. Deduce that if f and g both satisfy the same nth order ODE $f^{(n)} = F(t, f, \ldots, f^{(n-1)})$ on I, where F satisfies the hypotheses of the Picard-Lindelöf theorem, and are equal on a set S that has an accumulation point in I then they are equal at every point of I. (This is an ODE analogue of the *identity principle* in complex analysis.)

3. Deduce that if I is a *closed, bounded* interval then any two solutions to the same nth order ODE of the form $f^{(n)} = F(t, f, \ldots, f^{(n-1)})$ defined on I, where F satisfies the hypotheses of the Picard-Lindelöf theorem, are either equal or coincide at at most finitely many points. Is this true for $I = \mathbb{R}$?

Exercise 34. Verify in detail that Theorem 3.1 (the Picard-Lindelöf Theorem) and Proposition 2.19 imply Corollary 3.11.

Exercise 35. Formulate and prove a precise version of the following statement: For every $1 \leq m \leq n$, every nth order ODE for a function f taking values in $\mathbb{R}^d$ is equivalent to an mth order ODE for a function g taking values in $\mathbb{R}^{(n-m+1)d}$.

3.4 Maximal solutions

The Picard-Lindeloöf theorem has the annoying feature that it only tells us about existence and uniqueness of solutions *locally*, in a small interval around our starting time t_0. We now discuss the theory of *maximal* solutions, where we extend our solution to be defined on as big an interval as possible.

We first need to introduce the appropriate "global" analogue of the Lipschitz condition in the Picard-Lindelöf Theorem. Let $\Omega \subseteq \mathbb{R}^n$ be a set. A function $F : \Omega \to \mathbb{R}^m$ is said to be **Lipschitz** if there exists a constant M such that $\|F(x) - F(y)\| \leq M\|x - y\|$ for every $x, y \in \Omega$. (Recall that we interpret this as a 'bounded slope' condition; we'll interpret it in terms of the derivative of F momentarily.) We say that $F : \Omega \to \mathbb{R}^m$ is **locally Lipschitz** if for each $x \in \Omega$ there exists $\varepsilon > 0$ so that the restriction of F to $\{y \in \Omega : \|y - x\| \leq \varepsilon\}$ is Lipschitz. (In particular, Lipschitz functions are always locally Lipschitz.)

Proposition 3.12. *If $\Omega \subseteq \mathbb{R}^n$ is open and $F : \Omega \to \mathbb{R}^m$ is continuously differentiable then F is locally Lipschitz. If the derivative of F is bounded then F is Lipschitz.*

Proof. Let $\Omega \subseteq \mathbb{R}^n$ be an open set and let $F : \Omega \to \mathbb{R}^m$ be continuously differentiable. For any two points $x, y \in \Omega$ with $x \leq y$ we have that

$$\|F(x) - F(y)\| = \left\| \int_0^1 \frac{\partial}{\partial t} F(x + t(y - x)) \, \mathrm{d}t \right\| \leq \int_0^1 \left\| \frac{\partial}{\partial t} F(x + t(y - x)) \right\| \mathrm{d}t.$$

Now, using the chain rule, we have that

$$\frac{\partial}{\partial t}F(x+t(y-x)) = \big[DF(x+t(y-x))\big]\,(y-x)$$

and hence that

$$\left\|\frac{\partial}{\partial t}F(x+t(y-x))\right\| \leq \|DF(x+t(y-x))\|_{\mathrm{op}}\|y-x\|.$$

If F has bounded derivative, so that $\|DF(x)\|_{\mathrm{op}} \leq M$ for every $t \in \Omega$ and some $M < \infty$, then

$$\|F(x) - F(y)\| \leq M\|x-y\|$$

as claimed. More generally, since Ω is open, for each $x \in \Omega$ there exists $\varepsilon > 0$ such that the closed ball $B = \{y : \|y-x\| \leq \varepsilon\}$ is contained in Ω. Since this set is closed and bounded and the restriction of DF to this set is continuous, the supremum $\sup_{y\in B}\|DF(y)\|_{\mathrm{op}}$ is finite, and we have as above that

$$\|F(y) - F(z)\| \leq \sup_{y\in B}\|DF(y)\|_{\mathrm{op}} \cdot \|y-z\|$$

for every $y, z \in B$. Since $x \in \Omega$ was arbitrary, F is locally Lipschitz as claimed. □

Example 3.13. The function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = |x|$ is Lipschitz but not continuously differentiable.

Example 3.14. The function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^2$ is locally Lipschitz but not Lipschitz.

Given an open set $\Omega \subseteq \mathbb{R}^{1+nd}$ and a function $F : \Omega \to \mathbb{R}^d$, where we think of the first coordinate as time, we say that F is **locally space-Lipschitz** if for every $(t, \mathbf{x}) \in \Omega$ there exists $\varepsilon > 0$ and $M < \infty$ such that $\|F(s,\mathbf{z}) - F(s,\mathbf{y})\| \leq M\|\mathbf{z}-\mathbf{y}\|$ for every $(s,\mathbf{y}), (s,\mathbf{z}) \in \Omega$ with $\|(s,\mathbf{y}) - (t,\mathbf{x})\|, \|(s,\mathbf{z}) - (t,\mathbf{x})\| \leq \varepsilon$. (This terminology is not standard.) Being locally space-Lipschitz is a weaker condition than being locally Lipschitz, so that every continuously differentiable function is locally space-Lipschitz.

We now have everything we need to discuss the "global" version of Picard-Lindelöf. We say that a solution $(\tilde{I}, \tilde{f})$ is an **extension** of a solution (I, f) if $I \subseteq \tilde{I}$ and the restriction of $\tilde{f}$ to I is equal to f. (In particular, every solution is an extension of itself.) We say that a solution (I, f) is **maximal** if it has no extensions other than itself.

Theorem 3.15 (Global Picard-Lindelöf)**.** *Let $\Omega \subseteq \mathbb{R}^{1+nd}$ be an open set and let $F : \Omega \to \mathbb{R}^d$ be a continuous, locally space-Lipschitz function. For each initial condition $(t_0, \mathbf{x}_0) \in \Omega$ there exists a unique maximal solution $(I_{\max}, f_{\max})$ to (IVP), which extends every other solution. Moreover, the interval $I_{\max}$ is open.*

We begin by proving that we can always "glue together" two solutions to make a solution defined on a bigger interval.

Lemma 3.16. *Let $\Omega \subseteq \mathbb{R}^{1+nd}$ be an open set, let $F : \Omega \to \mathbb{R}^d$ be a continuous, locally space-Lipschitz function, and let $(t_0, \mathbf{x}_0) \in \Omega$. If (I_1, f_1) and (I_2, f_2) are two solutions to (IVP) then f_1 and f_2 coincide on $I_1 \cap I_2$ and if we define a function $f : I_1 \cup I_2 \to \mathbb{R}^d$ by*

$$f(t) = \begin{cases} f_1(t) & t \in I_1 \\ f_2(t) & t \in I_2 \end{cases} \tag{3.7}$$

then f is a solution to (IVP) that extends both (I_1, f_1) and (I_2, f_2).

Remark 3.17. A pretentious way to state this lemma is that the set of solutions to a given (continuous, locally space-Lipschitz) ODE form a "sheaf".

Proof of Lemma 3.16. Since F is locally space-Lipschitz, there exists a non-empty open set $U \subseteq \Omega$ containing the point (t_0, x_0) such that the restriction of $F(t, \mathbf{x})$ to U is Lipschitz in $\mathbf{x}$. As such, the Picard-Lindelöf Theorem implies that there exists $\varepsilon > 0$ such that (IVP) has a *unique* solution on every non-trivial closed interval of length at most ε that contains t_0, and hence that the two solutions f_1 and f_2 coincide on $I_1 \cap I_2 \cap [t_0 - \varepsilon/2, t_0 + \varepsilon/2]$. Let

$$\begin{aligned} t_+ &= \inf\{t \geq t_0 : f_1(t) \neq f_2(t) \text{ or } t \notin I_1 \cap I_2\} \\ t_- &= \sup\{t \leq t_0 : f_1(t) \neq f_2(t) \text{ or } t \notin I_1 \cap I_2\}. \end{aligned}$$

We want to prove that $t_+ = \sup I_1 \cap I_2$ and that $t_- = \inf I_1 \cap I_2$: This implies that f_1 and f_2 coincide on $I_1 \cap I_2$ since they are both continuous.

We will prove that $t_+ = \sup I_1 \cap I_2$, the proof of the claim about t_- being similar. Suppose not, so that $t_+ < \sup I_1 \cap I_2$. Since f_1 and f_2 are n-times differentiable, their first $n-1$ derivatives are continuous and $(t_+, f_1(t_+), \ldots, f_1^{(n-1)}(t_+)) = (t_+, f_2(t_+), \ldots, f_2^{(n-1)}(t_+)) \in \Omega$. Since F is locally space-Lipschitz on Ω there exists a non-empty open set U' around this point so that F is Lipschitz on U', and it follows by Picard-Lindelöf that there exists $\varepsilon > 0$ such that f_1 and f_2 coincide on $[t_+ - \varepsilon, t_+ + \varepsilon]$. This contradicts the definition of t_+, so that in fact f_1 and f_2 coincide on $I_1 \cap I_2$ as claimed.

It remains to verify that f is differentiable and satisfies (IVP). In general, if we glue together two functions f_1 and f_2 that are defined on two non-trivial intervals I_1 and I_2 of non-empty intersection $I_1 \cap I_2$ and coincide on this interval as in (3.7), the only way for the resulting function to not be n-times differentiable is for I_1 and I_2 to intersect at a single point and for f to have distinct left and right mth derivatives at this point for some $1 \leq m \leq n$. In our case this cannot happen since the intervals I_1 and I_2 must both contain t_0 and solutions to (IVP) have specified mth derivatives at this point for every $1 \leq m \leq n$. Thus, f is n-times differentiable and satisfies (IVP) since f_1 and f_2 both do. □

Proof of Global Picard-Lindelöf. Since F is locally space-Lipschitz, there exists a non-empty open set $U \subseteq \mathbb{R}^{1+nd}$ containing the point $(t_0, \mathbf{x}_0)$ such that the restriction of $F(t, \mathbf{x})$ to U is Lipschitz in $\mathbf{x}$. It follows from Picard-Lindelöf that there is at least one solution (I, f) of (IVP). To define a maximal solution, we first take the maximal interval

$$I_{\max} = \bigcup \big\{ I : (I, f) \text{ is a solution of (IVP)} \big\}.$$

(Note that if a unique maximal solution exists then its domain *must* be given by this expression!) We define a function $f_{\max} : I_{\max} \to \mathbb{R}^d$ by, for each $x \in I_{\max}$, picking a solution (I, f) to (IVP) with $x \in I$, one of which must exist by definition of $I_{\max}$, and taking $f_{\max}(x) = f(x)$ – Lemma 3.16 implies that the choice of solution (I, f) does not affect the value of $f_{\max}(x)$! For each $x \in I_{\max}$ and each solution (I, f) to (IVP) with $x \in I$ we have moreover that the restriction of $f_{\max}$ coincides with f, and it follows that $f_{\max}$ is n-times differentiable and is a solution to (IVP). Since $(I_{\max}, f_{\max})$ extends every solution to (IVP), it is the unique maximal solution to (IVP).

We now prove that $I_{\max}$ is open. If $I_{\max}$ is *not* open, then either $\sup I_{\max} < \infty$ and $\sup I_{\max} \in I$ or $\inf I_{\max} > -\infty$ and $\inf I_{\max} \in I$ (or both). We will rule out the case that $\sup I_{\max} \in I_{\max}$, the case that $\inf I_{\max} \in I_{\max}$ being similar. Suppose that (I, f) is a solution to (IVP) with $\sup I \in I$. Write $\sup I = b$. Since F is locally Lipschitz, there exists a non-empty open set $U' \subseteq \Omega$ containing the point $(b, f(b), \dots, f^{(n-1)}(b))$ such that the restriction of $F(t, \mathbf{x})$ to U' is Lipschitz in $\mathbf{x}$. As such, the Picard-Lindelöf theorem implies that there exists $\varepsilon > 0$ and a solution g to the ODE $g^{(n)} = F(t, g, \dots, g^{(n-1)}(t))$ defined on $(b - \varepsilon, b + \varepsilon)$ satisfying $(g(b), g'(b), \dots, g^{(n-1)}(b)) = (f(b), f'(b), \dots, f^{(n-1)}(b))$. Thus, if we let $\tilde{I} = I \cup (b - \varepsilon, b + \varepsilon)$ then the function $\tilde{f} : \tilde{I} \to \mathbb{R}^d$ defined by

$$\tilde{f}(x) = \begin{cases} f(x) & x \in I \\ g(x) & x \in \tilde{I} \setminus I \end{cases}$$

in n-times differentiable and satisfies the (IVP). It follows that any solution (I, f) to (IVP) such that $\sup I \in I$ admits a non-trivial extension (i.e., an extension that is not equal to (I, f) itself), and hence that $\sup I_{\max} \notin I_{\max}$. □

Remark 3.18. If we consider an ODE rather than an IVP, the Global Picard-Lindelöf theorem implies that if F is locally space-Lipschitz then there is exactly one maximal solution passing through each phase-space point $(t_0, \mathbf{x}_0)$, which extends every other solution passing through this point. Indeed, "passing through this point" is equivalent to solving an appropriate IVP!

Exercise 36. Let $\Omega \subseteq \mathbb{R}^{1+nd}$ be an open set, let $F : \Omega \to \mathbb{R}^d$ be a continuous, locally space-Lipschitz function, let $(t_0, \mathbf{x}_0) \in \Omega$ and let (I, f) be a maximal solution to (IVP). Prove that there must exist a sequence (t_k) in I with $t_k \to \sup I$ such that $(t_k, f(t_k), f'(t_k), \dots, f^{(n-1)}(t_k))$ either has a coordinate converging to infinity or converges to a point that does not belong to Ω. In other words, a solution can only fail to be extended if it blows up or leaves the domain

of definition of the ODE.

3.5 Autonomous equations

Before moving on, let us briefly discuss *autonomous* equations. As we briefly mentioned back at the beginning of the course, an **autonomous** ODE is one of the form

$$f^{(n)} = F(f, f', \dots, f^{(n-1)}),$$

i.e., where the right hand side does not directly depend on the dependent variable t. An important property of autonomous ODEs is that if the function $t \mapsto f(t)$ solves an autonomous ODE then so does the 'time-shifted' function $t \mapsto f(t - t_0)$ for every $t_0 \in \mathbb{R}$.

When the hypotheses of Global Picard-Lindelöf are satisfied, there is an easy way to check that we've found all solutions to an autonomous ODE:

Theorem 3.19 (Solving autonomous equations by shifting)**.** *Let $\Omega \subseteq \mathbb{R}^{nd}$ be open, let $F : \Omega \to \mathbb{R}^d$ be locally space-Lipschitz, and suppose that S is a set of maximal solutions to the equation $f^{(n)} = F(f, \dots, f')$. If for each $x_0 = (x_{0,0}, \dots, x_{0,n}) \in \Omega$ there exists $(I, f) \in S$ and $t \in I$ such that $(f(t), \dots, f^{(n-1)}) = x_0$ then every maximal solution to $f^{(n)} = F(f, \dots, f^{(n-1)})$ is of the form $(I + t_0, f(t - t_0))$ for some $t_0 \in \mathbb{R}$ and $(I, f) \in S$ where the shifted interval $I + t_0$ is defined by $I + t_0 = \{t + t_0 : t \in I\}$.*

In other words, if we have enough maximal solutions to visit every point of the phase space in which the autonomous ODE $f^{(n)} = F(f, \dots, f^{(n-1)})$ is defined, then every other maximal solution can be found by 'time-shifting' one of these solutions.

Proof. First note that if $(I, f(t))$ is a maximal solution to the ODE then so is $(I + t_0, f(t - t_0))$ for each $t_0 \in \mathbb{R}$: That it is a solution follows since the mth derivative of the time-shift of a function is always equal to the time-shift of the mth derivative of that function. That it is maximal follows since if it weren't then we could define a non-trivial extension of $(I, f(t))$ by time-shifting back the non-trivial extension of $(I + t_0, f(t - t_0))$. The fact that *every* maximal solution is a time-shift of a solution in S follows by Global Picard-Lindelöf (applied to the function $\tilde{F} : \mathbb{R} \times \Omega$ defined by $\tilde{F}(t, x) = F(x)$) since this set of maximal solutions includes one going through each possible initial condition $(t_0, x_0) \in \mathbb{R} \times \Omega$. □

In particular, if we have a solution $f : \mathbb{R} \to \mathbb{R}$ to a first-order, one-dimensional autonomous ODE $f' = F(f)$ that F is locally space-Lipschitz on $\mathbb{R}$ and f is a bijection, then every maximal solution to this equation is of the form $(\mathbb{R}, f(t - t_0))$ for some $t_0 \in \mathbb{R}$.

3.6 Smoothness of solutions

Part of our definition of what it meant for a function to be a solution of an nth order ODE was for it to be n-times differentiable. In fact, we can often guarantee from first principles that solutions to ODEs are smooth:

Proposition 3.20. *Let $\Omega \subseteq \mathbb{R}^{1+nd}$ be open and let $F : \Omega \to \mathbb{R}^d$ be k-times differentiable. If (I, f) is a solution to the ODE $f^{(n)} = F(t, f, \dots, f^{(n-1)})$ then f is $k+n$-times differentiable. In particular, if F is smooth then f is smooth.*

Proof. We will prove that f is $m+n$-times differentiable for every $0 \leq m \leq k$ by induction on m. The base case $m = 0$ holds by assumption. Let $0 < m \leq k$ and suppose that f is $(n+m-1)$-times differentiable, in which case the function mapping t to $(t, f, \dots, f^{(n-1)})$ is m times differentiable. The (multivariable) chain rule implies that the composition of two m-times differentiable functions is differentiable, and since F is m-times differentiable by assumption it follows that $F(t, f, \dots, f^{(n-1)})$ is m-times differentiable. This completes the proof of the induction step since $f^{(n)} = F(t, f, \dots, f^{(n-1)})$. □

Exercise 37. Give an example where F is Lipschitz but not differentiable, and the solution to the first-order autonomous ODE $f' = F(f)$ is not twice-differentiable.

Later we will prove that solutions to ODEs are often *real analytic*, a much stronger condition than being smooth.

3.7 Dependence of solutions on coefficients and initial conditions

We now study the dependence of solutions to ODEs on their initial conditions and coefficients. We will not prove the most general version of these theorem that we possibly could; there are also versions that only require F to be locally space-Lipschitz. The version that we now state is sufficiently general to handle all linear ODEs (even if they have non-constant coefficients). Note that the part of the theorem that rules out finite-time blow-up (unless the ODE becomes undefined) really is specific to the case that F is space-Lipschitz rather than locally space-Lipschitz.

Proposition 3.21 (Continuity of solutions as functions of their initial conditions)**.** *Suppose that $I \subseteq \mathbb{R}$ is a non-trivial open interval and that $F : I \times \mathbb{R}^{nd} \to \mathbb{R}^d$ is a continuous function such that for each closed, bounded interval $J \subseteq I$, the restriction of F to J is space-Lipschitz in the sense that there exists $M(J) < \infty$ such that $\|F(t, \mathbf{x}) - F(t, \mathbf{y})\| \leq M\|\mathbf{x} - \mathbf{y}\|$ for every $t \in \mathbb{R}$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{nd}$. Then every maximal solution to the ODE $f^{(n)} = F(t, f, \dots, f^{(n-1)})$ has domain I. Moreover, if $J \subseteq I$ is a closed, bounded interval containing the point t_0 and (I, f) and (I, g) are the unique solutions passing through the points $(t_0, \mathbf{x}_0)$ and $(t_0, \mathbf{y}_0)$ for some $\mathbf{x}_0$ and $\mathbf{y}_0$ in $\mathbb{R}^{nd}$ then*

$$\|\mathbf{f}(t) - \mathbf{g}(t)\| \leq e^{M(J)|t-t_0|}\|\mathbf{x}_0 - \mathbf{y}_0\|$$

for every $t \in J$.

In other words, solutions to space-Lipschitz ODEs that start within distance ε of each other remain close until time at least of order $\log 1/\varepsilon$.

Example 3.22. The ODE $f' = f$ is Lipschitz with constant $M = 1$. The solutions to this ODE starting at 0 and ε at time 0 are 0 and εe^t respectively, and show that the inequality of Proposition 3.21 cannot be improved.

This proposition has the following corollary.

Corollary 3.23 (No finite-time blow-up for linear ODEs)**.** *Let I be an open interval and suppose that $a_{n-1}, \dots, a_0, b : I \to \mathbb{R}$ are continuous. Then every $\mathbf{x}_0 \in \mathbb{R}^n$ and $t_0 \in I$, there is a unique maximal solution to the ODE $f^{(n)} + a_{n-1} f^{(n-1)} + \cdots a_1 f' + a_0 f = b$ with $\mathbf{f}(t_0) = \mathbf{x}_0$, and this maximal solution has domain I.*

We will prove this proposition using *Grönwall's inequality*, an elementary lemma that is very useful throughout the study of ODEs.

Lemma 3.24 (Grönwall's inequality)**.** *Let I be a non-trivial interval, let $t_0 \in I$, and suppose that $f, g, h : I \to \mathbb{R}$ are continuous functions such that f is differentiable and*

$$f'(t) \leq h(t) + g(t) f(t)$$

for every $t \geq t_0$ in I. Then

$$f(t) \leq \exp\left[\int_{t_0}^{t} g(s)\,\mathrm{d}s\right]\left(f(t_0) + \int_{t_0}^{t} \exp\left[-\int_{t_0}^{s} g(x)\,\mathrm{d}x\right] h(s)\,\mathrm{d}s\right)$$

for every $t \geq t_0$ in I. In other words, functions satisfying the differential inequality $f'(t) \leq h(t) + g(t)f(t)$ *are bounded by solutions to the differential* equation $f'(t) = h(t) + g(t)f(t)$*.*

Proof. Define $u(t) = \exp\left[-\int_{t_0}^{t} g(t)\right]$. Then we have by the chain rule, product rule, and fundamental theorem of calculus that

$$(fu)' = f'u - gfu = u(f' - gf) \leq uh$$

where we used the inequality $f' \leq h + fg$ (together with the fact that u is non-negative) in the final inequality. It follows by the fundamental theorem of calculus that

$$f(t)u(t) - f(t_0)u(t_0) = \int_{t_0}^{t} (fu)'\,\mathrm{d}s \leq \int_{t_0}^{t} u(s)h(s)\,\mathrm{d}s,$$

which is equivalent to the claim since $u(t_0) = 1$. □

Unfortunately, the main thing we would like to "apply" Grönwall to is $\|f\|$, which is not differentiable. As such, it will be useful to have an "integral" version of Grönwall that does not require differentiability of f. (On the other hand, this integral version will require that g is non-negative, which was not needed for the differential version.)

Lemma 3.25 (Grönwall's inequality, integral form)**.** *Let I be a non-trivial interval, let $t_0 \in I$, and suppose that $f, g : I \to \mathbb{R}$ are continuous functions such g is non-negative and*

$$f(t) - f(t_0) \leq \int_{t_0}^{t} h(s) + g(s)f(s)\,\mathrm{d}s$$

for every $t \geq t_0$ in I. Then

$$f(t) \leq \exp\left[\int_{t_0}^{t} g(s)\,\mathrm{d}s\right]\left(f(t_0) + \int_{t_0}^{t} \exp\left[-\int_{t_0}^{r} g(x)\,\mathrm{d}x\right] h(s)\,\mathrm{d}s\right)$$

for every $t \geq t_0$ in I.

Proof. Consider the function $v(t)$ defined by

$$v(t) = \exp\left[-\int_{t_0}^{t} g(s)\,\mathrm{d}s\right]\int_{t_0}^{t} g(s)f(s)\,\mathrm{d}s.$$

We can differentiate v using the product rule, chain rule, and fundamental theorem of calculus to obtain that

$$\begin{aligned} v'(t) &= -g(t)\exp\left[-\int_{t_0}^{t} g(s)\,\mathrm{d}s\right]\int_{t_0}^{t} g(s)f(s)\,\mathrm{d}s + g(t)f(t)\exp\left[-\int_{t_0}^{t} g(s)\,\mathrm{d}s\right] \\ &= g(t)\exp\left[-\int_{t_0}^{t} g(s)\,\mathrm{d}s\right]\cdot\left(f(t) - \int_{t_0}^{t} g(s)f(s)\,\mathrm{d}s\right). \end{aligned}$$

Using the assumption that g is non-negative, we obtain that

$$v'(t) \leq g(t)\exp\left[-\int_{t_0}^{t} g(s)\,\mathrm{d}s\right]\left(f(t_0) + \int_{t_0}^{t} h(s)\,\mathrm{d}s\right)$$

and hence that

$$\begin{aligned} v(t) = v(t) - v(t_0) &\leq \int_{t_0}^{t} g(r)\exp\left[-\int_{t_0}^{r} g(s)\,\mathrm{d}s\right]\left(f(t_0) + \int_{t_0}^{r} h(s)\,\mathrm{d}s\right)\mathrm{d}r \\ &= -\exp\left[-\int_{t_0}^{t} g(s)\,\mathrm{d}s\right]\left(f(t_0) + \int_{t_0}^{t} h(s)\,\mathrm{d}s\right) + f(t_0) + \int_{t_0}^{t}\exp\left[-\int_{t_0}^{r} g(s)\,\mathrm{d}s\right] h(r)\,\mathrm{d}r, \end{aligned}$$

where we used integration by parts in the second line. It follows by definition of $v(t)$ that

$$\begin{aligned} \int_{t_0}^{t} g(s)f(s)\,\mathrm{d}s = v(t)\exp\left[\int_{t_0}^{t} g(s)\,\mathrm{d}s\right] &\leq -\int_{t_0}^{t} h(s)\,\mathrm{d}s - f(t_0) \\ &\quad + \exp\left[\int_{t_0}^{t} g(s)\,\mathrm{d}s\right] f(t_0) + \exp\left[\int_{t_0}^{t} g(s)\,\mathrm{d}s\right]\int_{t_0}^{t}\exp\left[-\int_{t_0}^{r} g(s)\,\mathrm{d}s\right] h(r)\,\mathrm{d}r \end{aligned}$$

and hence that

$$f(t) \leq \exp\left[\int_{t_0}^{t} g(s)\,\mathrm{d}s\right]\left(f(t_0) + \int_{t_0}^{t} \exp\left[-\int_{t_0}^{r} g(s)\,\mathrm{d}s\right] h(r)\,\mathrm{d}r\right)$$

as claimed. □

Let us now prove Proposition 3.21.

Proof of Proposition 3.21. We first prove the claim that maximal solutions are defined on I. It suffices to prove that if $t_0 \in I$ then every maximal solution defined on an interval containing t_0 has domain containing J for every closed, bounded interval contained in I and containing t_0. Fix t_0 and J and let $M = M(J)$. Let (I', f) be a solution to the ODE with $t_0 \in I'$ and let $\mathbf{f} = (f, f', \ldots, f^{(n-1)})$ be the corresponding phase space solution. Then

$$\|\mathbf{f}(t) - \mathbf{f}(t_0)\| = \|\int_{t_0}^{t} \mathbf{F}(s, \mathbf{f}(s))\,\mathrm{d}s\| \leq \left|\int_{t_0}^{t} \|\mathbf{F}(s, \mathbf{f}(s))\|\,\mathrm{d}s\right|$$

where $\mathbf{F}$ is the phase-space version of F. It follows from the triangle inequality that $\mathbf{F}$ is space-Lipschitz whenever F is, with the same constant M. As such, we can bound

$$\|\mathbf{F}(s, \mathbf{f}(s))\| \leq \|\mathbf{F}(s, \mathbf{f}(t_0))\| + M\|\mathbf{f}(t) - \mathbf{f}(t_0)\|$$

for every $t \in I' \cap J$. As such, we can apply the integral form of Grönwall's inequality (with $h(t) = \|\mathbf{F}(t, \mathbf{f}(t_0))\|$ and $g(t) \equiv M$) to deduce that

$$\|\mathbf{f}(t) - \mathbf{f}(t_0)\| \leq e^{M(t-t_0)} \int_{t_0}^{t} e^{-M(s-t_0)} \|\mathbf{F}(s, \mathbf{f}(t_0))\|\,\mathrm{d}s$$

for every $t \geq t_0$ in $I' \cap J$. Applying the same argument to $f(-t)$ yields more generally that

$$\|\mathbf{f}(t) - \mathbf{f}(t_0)\| \leq e^{M|t-t_0|} \left|\int_{t_0}^{t} e^{-M|s-t_0|} \|\mathbf{F}(s, \mathbf{f}(t_0))\|\,\mathrm{d}s\right| \tag{3.8}$$

for every $t \in I$. The only important feature of this inequality for our current purposes is that $\|\mathbf{f}(t)\|$ is bounded by a continuous function that is defined on all of I, and hence is bounded on J. Since there is no special role for t_0 in this inequality, we obtain more generally that

$$\|\mathbf{f}(t_2) - \mathbf{f}(t_1)\| \leq e^{M|t_2-t_1|} \left|\int_{t_1}^{t_2} e^{-M|s-t_1|} \|\mathbf{F}(s, \mathbf{f}(t_1))\|\,\mathrm{d}s\right| \tag{3.9}$$

for every $t_1, t_2 \in J$.

To use this inequality, we will need the following elementary fact.

Exercise 38. Let $(V, \|\cdot\|)$ be a normed vector space and let $a < b$ be real numbers. If $f : (a, b) \to V$ is n-times continuously differentiable and the limits $\lim_{t\uparrow b} f(t)$ and $\lim_{t\uparrow b} f^{(m)}(t)$ exist for every $1 \leq m \leq n$, then the extension of f to $(a, b]$ defined by taking $f(b) = \lim_{t\uparrow b} f(t)$ is n-times continuously differentiable.

We will prove that if (I', f) is maximal then $\sup J \in I'$, the proof that $\inf J \in I'$ being similar. It suffices to prove that if $\sup J \notin I'$ then the solution (I', f) is not maximal. Suppose to this end that $\sup J \notin I'$, and let $t_+ = \sup I' \leq \sup J$. It follows from (3.9) that $\|\mathbf{f}(t)\|$ is bounded on $[t_0, t_+]$, and since $\mathbf{F}$ is continuous on $I \times \mathbb{R}^{nd}$ it follows by the extreme value theorem that $\|\mathbf{F}(s, \mathbf{f}(t))\|$ is bounded on the set $\{(s, t) : t_0 \leq s, t < t_+\}$. Letting C be the maximum of $\|\mathbf{F}(s, \mathbf{f}(t))\|$ on this set, it follows from (3.9) that

$$\|\mathbf{f}(t_2) - \mathbf{f}(t_1)\| \leq Ce^{M|t_2-t_1|} \left| \int_{t_1}^{t_2} e^{-M|s-t_1|} \, \mathrm{d}s \right| \leq Ce^{M|t_2-t_1|}|t_2 - t_1|$$

for every $t_0 \leq t_1 \leq t_2 < \sup I$. Since the right hand side is small when $|t_2-t_1|$ is small, it follows that the limit $\lim_{t\uparrow t_+} \mathbf{f}(t)$ is well-defined. Since F is continuous, $\lim_{t\uparrow t_+} f^{(n)} = \lim_{t\uparrow t_+} F(t, \mathbf{f}(t))$ is also well-defined and equal to $F(t_+, \lim_{t\uparrow t_+} \mathbf{f}(t))$. As such, we can extend f to the interval $I' \cup \{t_+\}$ by setting $f(t_+) = \lim_{t\uparrow t_+} f(t)$, which yields a solution to the ODE defined on this larger interval by the above exercise. This shows that (I', f) was not maximal, completing the proof.

It remains to prove the claim concerning stability with respect to initial conditions. Let (I, f) and (I, g) be the two maximal solutions passing through the points $(t_0, \mathbf{x}_0)$ and $(t_0, \mathbf{y}_0)$ respectively, let $\mathbf{f}$ and $\mathbf{g}$ be the corresponding phase-space solutions, and let J be a closed, bounded interval contained in I and containing t_0. Writing $M = M(J)$, we have that

$$\begin{aligned} \|\mathbf{f}(t) - \mathbf{g}(t)\| &= \left\| \int_{t_0}^{t} \mathbf{F}(s, \mathbf{f}(s)) - \mathbf{F}(s, \mathbf{g}(s)) \, \mathrm{d}s \right\| \\ &\leq \left| \int_{t_0}^{t} \left\| \mathbf{F}(s, f(s)) - \mathbf{F}(s, g(s)) \right\| \mathrm{d}s \right| \leq M \left| \int_{t_0}^{t} \|\mathbf{f}(s) - \mathbf{g}(s)\| \, \mathrm{d}s \right| \end{aligned}$$

for every $t \in J$, and applying the integral form of Grönwall yields that

$$\|\mathbf{f}(t) - \mathbf{g}(t)\| \leq e^{M|t-t_0|} \|\mathbf{f}(t_0) - \mathbf{g}(t_0)\|$$

for all $t \in J$ with $t \geq t_0$. An analogous inequality holds also for $t \leq t_0$ by similar reasoning. This is stronger than the claimed inequality. □

We stress again that the proposition we have just proved is just one particularly simple instance of a "continuity of the solution as a function of the initial condition" result; one can also prove similar theorems under the weaker assumption that the function is only locally space-Lipschitz.

We next deduce a similar theorem about continuous dependence of solutions on the function F_a. Again, the version we are stating here is *not* the most general theorem you could prove to this effect.

Proposition 3.26 (Continuity of solutions as functions of coefficients)**.** *Suppose that $I \subseteq \mathbb{R}$ is a non-trivial open interval, and that for each $a \in \mathbb{R}^m$ we have a continuous function $F_a : I \times \mathbb{R}^{nd} \to \mathbb{R}^d$. Suppose further that for each closed bounded interval $J \subseteq I$ there exists $M(J) < \infty$ such that $\|F_a(t, \mathbf{x}) - F_b(t, \mathbf{y})\| \leq M(J)\|(\mathbf{x}, a) - (\mathbf{y}, b)\|$ for every $t \in J$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{nd}$, and $a, b \in \mathbb{R}^m$. If (I, f) and (I, g) are the unique maximal solutions to the ODEs $f^{(n)} = F_a(t, f, \ldots, f^{(n-1)})$ and $g^{(n)} = F_b(t, g, \ldots, g^{(n-1)})$ passing through the point $(t_0, \mathbf{x}_0)$ for some $\mathbf{x}_0$ in $\mathbb{R}^{nd}$ and $a, b \in \mathbb{R}^m$ then*

$$\|\mathbf{f}(t) - \mathbf{g}(t)\| \leq e^{M|t-t_0|}\|a - b\|$$

for every $t \in \mathbb{R}$.

Proof. Consider the first-order, $(nd+m)$-dimensional ODE $(\mathbf{f}, a)' = (\mathbf{F}_a(t, \mathbf{f}), 0)$, where $\mathbf{F}_a$ is the phase-space version of F_a. The solutions to this ODE have a constant value of their second coordinate a, and are in bijection with the solutions to the original ODE; the constant second coordinate encodes the parameter a in the function F_a. As such, the proposition follows from Proposition 3.21. □

Remark 3.27. A different way to prove stability under changing coefficients and initial conditions is using the *Arzela-Ascoli theorem*, which gives conditions under which a sequence of continuous functions has subsequential limits in the uniform norm. Using this theorem, one shows that if one has a convergent sequence of initial conditions and a convergent sequence of functions encoding the ODE, then the solutions to the relevant IVP must have a subsequential limit, and that any such subsequential limit must be a solution to the limiting IVP. Under the hypotheses of Picard-Lindelöf this solution is unique, so that in fact the solutions to the IVPs in the sequence really do converge to the solution of the limiting IVP. One complication in doing this is that one must define all relevant "spaces of solutions" in a way that accounts for the fact that different solutions might have different domains, even when these solutions are maximal. Note that this proof method is more general than ours, but does not provide quantitative estimates on how close the two solutions must be to each other.

3.8 Inhomogeneous linear ODEs and Duhamel's principle

Let us now consider the case of *inhomogeneous* linear ODEs

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

for some functions $a_{n-1}, \ldots, a_0$ and b. As before, we can write this equation in the form

$$\mathbf{f}' = A\mathbf{f} + \mathbf{b}$$

Recall that if $\mathbf{f}$ is any (phase-space) solution to this ODE, then every other solution can be written in the form $\mathbf{f} + \mathbf{g}$, where $\mathbf{g}$ is a solution to the homogeneous ODE $\mathbf{g}' = A\mathbf{g}$. As such, to solve inhomogeneous linear ODEs we just need to be able to solve the corresponding homogenenous linear ODE and be able to find *one* solution to the inhomogeneous ODE. This is often called finding a **particular solution**. In practice, the best way to do this is often "by inspection", i.e. just guessing a particular solution based on your experience solving ODEs. In this section we will discuss a systematic way to find a particular solution known as *Duhamel's principle*.

Before discussing Duhamel's principle, let us mention that is is usually possible to think of inhomogeneous linear ODEs as a kind of *homogeneous* linear ODE in one higher dimension, provided that $\log b$ is well-defined and differentiable. For example, if b is constant then

$$\begin{pmatrix} f^{(n-1)} \\ f^{(n-2)} \\ \vdots \\ f' \\ f \\ b \end{pmatrix}' = \begin{pmatrix} b - a_{n-1}f^{(n-1)} - \cdots - a_1 f' - a_0 f \\ f^{(n-1)} \\ \vdots \\ f'' \\ f' \\ 0 \end{pmatrix}$$

$$= \begin{pmatrix} -a_{n-1} & -a_{n-2} & \cdots & -a_2 & -a_1 & -a_0 & 1 \\ 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & 0 \\ 0 & 0 & \cdots & 0 & 1 & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} f^{(n-1)} \\ f^{(n-2)} \\ \vdots \\ f' \\ f \\ b \end{pmatrix},$$

while if b is continuously differentiable and does not take the value zero then

$$\begin{pmatrix} f^{(n-1)} \\ f^{(n-2)} \\ \vdots \\ f' \\ f \\ b \end{pmatrix}' = \begin{pmatrix} b - a_{n-1}f^{(n-1)} - \cdots - a_1 f' - a_0 f \\ f^{(n-1)} \\ \vdots \\ f'' \\ f' \\ 0 \end{pmatrix}$$

$$= \begin{pmatrix} -a_{n-1} & -a_{n-2} & \cdots & -a_2 & -a_1 & -a_0 & 1 \\ 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & 0 \\ 0 & 0 & \cdots & 0 & 1 & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & (\log b)' \end{pmatrix} \begin{pmatrix} f^{(n-1)} \\ f^{(n-2)} \\ \vdots \\ f' \\ f \\ b \end{pmatrix}.$$

This means that we can always solve *constant coefficient* linear ODEs using matrix exponentiation, even if they are inhomogeneous. (Be careful to note however that the relevant matrices are not quite of the same form that we considered in the homogeneous case.)

We now discuss Duhamel's principle.

Proposition 3.28 (Duhamel's principle)**.** *Consider the linear ODE* $\mathbf{f}' = A\mathbf{f} + \mathbf{b}$*, where* A *and* b *depend continuously on* t *and are defined on some non-empty open interval* $I \subseteq \mathbb{R}$*. For each* $s \in I$*, let* f_s *be the solution to the homogeneous linear ODE* $\mathbf{f}' = A\mathbf{f}$ *with* $f_s^{(n-1)}(s) = b(s)$ *and* $f_s^{(m)}(s) = 0$ *for every* $m < n-1$*. For each* $t_0 \in I$*, the function defined on* I *by*

$$f(t) = \int_{t_0}^{t} f_s(t)\,\mathrm{d}s$$

is a solution to $\mathbf{f}' = A\mathbf{f} + \mathbf{b}$*. Consequently, every solution to this ODE can be expressed as the sum of this function and a solution to the homogeneous linear ODE* $\mathbf{f}' = A\mathbf{f}$*.*

The proof of Duhamel will use the following lemma, which is a consequence of Proposition 3.21; the proof is left as an exercise. This lemma also ensures that the function f from the statement of Duhamel is well-defined.

Lemma 3.29 (Regularity of solutions)**.** *For each* $0 \leq m \leq n$ *the function* $f_s^{(m)}(t) : I \times \mathbb{R}^n \to \mathbb{R}^n$ *depends continuously on* (s,t)*.*

We will also need the following standard fact about differentiating under the integral sign.

Theorem 3.30 (Differentiating under the integral sign)**.** *Let* $x_0 \leq x_1$*, let* $a : [x_0, x_1] \to \mathbb{R}$ *and* $b : [x_0, x_1] \to \mathbb{R}$ *be continuously differentiable with* $a(x) \leq b(x)$ *for every* $x_0 \leq x \leq x_1$*, and suppose that* $f(x,t)$ *is a function defined on some subset* Ω *of* $\mathbb{R}^2$ *such that* f *and its partial derivative* $\frac{\partial f}{\partial x}$ *are well-defined and continuous and* $\{(x,t) : x_0 \leq x \leq x_1, a(x) \leq t \leq b(x)\} \subseteq \Omega$*. Then* $\int_{a(x)}^{b(x)} f(x,t)\,\mathrm{d}t$ *is differentiable with derivative*

$$\frac{d}{dx}\int_{a(x)}^{b(x)} f(x,t)\,\mathrm{d}t = f(x,b(x))\frac{d}{dx}b(x) - f(x,a(x))\frac{d}{dx}a(x) + \int_{a(x)}^{b(x)} \frac{\partial}{\partial x} f(x,t)\,\mathrm{d}t.$$

Proof of Duhamel's principle. That fact ensures that we have the conditions we need to differentiate under the integral sign, so that

$$f' = \frac{d}{dt}\int_{t_0}^{t} f_s(t)\,\mathrm{d}s = f_t(t) + \int_{t_0}^{t} \frac{\partial}{\partial t} f_s(t)\,\mathrm{d}s = \int_{t_0}^{t} f_s'(t)\,\mathrm{d}s,$$

and by induction that

$$f^{(m)} = \frac{d^m}{dt^m}\int_{t_0}^{t} f_s(t)\,\mathrm{d}s = f_t^{(m-1)}(t) + \int_{t_0}^{t} f_s^{(m)}(t)\,\mathrm{d}s = \begin{cases} \int_{t_0}^{t} f_s^{(m)}(t)\,\mathrm{d}s & m < n \\ b(t) + \int_{t_0}^{t} f_s^{(m)}(t)\,\mathrm{d}s & m = n \end{cases}$$

for every $1 \leq m \leq n$. It follows by linearity of the integral that

$$\begin{aligned} f^{(n)}(t) + \sum_{i=0}^{n-1} a_i f^{(m)}(t) &= b(t) + \int_{t_0}^{t} f_s^{(n)} \, \mathrm{d}s + \sum_{i=0}^{m} a_i \int_{t_0}^{t} f_s^{(m)} \, \mathrm{d}s \\ &= b(t) + \int_{t_0}^{t} \left(f_s^{(n)} + \sum_{i=0}^{m} a_i f_s^{(m)} \right) \mathrm{d}s = b(t) \end{aligned}$$

as required. □

Let us now use Duhamel's principle to give the general solution to first-order linear ODEs.

Theorem 3.31 (First-order linear ODEs)**.** *Let $I \subseteq \mathbb{R}$ be a non-trivial open interval, let $t_0 \in I$, and let $a, b : I \to \mathbb{R}$ be continuous. Then every maximal solution to the first-order inhomogeneous linear ODE*

$$f' + af = b$$

is of the form

$$\left(I, e^{-\int_{t_0}^{t} a(s) \, \mathrm{d}s} \left(\int_{t_0}^{t} e^{\int_{t_0}^{s} a(u) \, \mathrm{d}u} b(s) \, \mathrm{d}s + C \right) \right).$$

Proof. We already know that every maximal solution has domain I. We also already know that every solution to the homogeneous linear ODE $f' = -af$ is of the form $f = Ce^{-\int_{t_0}^{t} a(s) \, \mathrm{d}s}$ for some $t_0 \in I$ and $C \in \mathbb{R}$. As such, the function f_s defined in the statement of Duhamel's principle can be written

$$f_s(t) = b(s) e^{-\int_{s}^{t} a(u) \, \mathrm{d}u},$$

so that

$$\int_{t_0}^{t} b(s) e^{-\int_{s}^{t} a(u) \, \mathrm{d}u} \, \mathrm{d}s = e^{-\int_{t_0}^{t} a(u) \, \mathrm{d}u} \int_{t_0}^{t} b(s) e^{\int_{t_0}^{s} a(u) \, \mathrm{d}u} \, \mathrm{d}s$$

in a solution to our inhomogeneous linear ODE. The claim follows since every solution can be written as the sum of this solution with some solution of the associated homogeneous linear ODE $f' = -af$. □

Exercise 39. In this exercise you will work through an alternative derivation of the solution to first-order inhomogeneous linear ODEs. Let $I \subseteq \mathbb{R}$ be a non-trivial open interval, let $t_0 \in I$, and let $a, b : I \to \mathbb{R}$ be continuous. Define the *integrating factor* $\mu(t) = e^{\int_{t_0}^{t} a(s) \, \mathrm{d}s}$. Show that if f solves the ODE $f' + af = b$ then $\mu(t) f(t)$ satisfies the ODE

$$(\mu f)' = \mu b.$$

Use this to give an alternative proof of Theorem 3.31.

Exercise 40. Find the maximal solution to the IVP

$$f' + \frac{3f}{t} = t^2 \qquad f(1) = 1/2.$$

Exercise 41. Find every maximal solution to the ODE $f' + tf = t^2$.

4 Separable equations

Now that we have laid the proper groundwork by establishing our main existence and uniqueness theorems, it's time to start solving some equations! We begin by studying **separable** equations, one of the most important and *easy to solve* classes of equations arising in applications. We say that a first-order ODE $f' = F(t, f(t))$ describing a one-dimensional function f is **separable** if F can be factored into a term depending only on t and a term depending only on $f(t)$ – in other words, a separable ODE is a one-dimensional, first-order ODE of the form

$$\frac{df}{dt} = F(t)G(f(t)).$$

Note that this includes the case that one of F or G is constant. Before stating any theorems, let us start by explaining how to solve such an ODE in heuristic terms. The (non-rigorous) idea is to imagine that we are free to 'rearrange' the equation to read

$$\frac{1}{G(f)}\,\mathrm{d}f = F(t)\,\mathrm{d}t$$

(of course this does not really make sense!) then integrate both sides to obtain that

$$\int \frac{1}{G(f)}\,\mathrm{d}f = \int F(t)\,\mathrm{d}t. \tag{4.1}$$

Of course, even if the method used to arrive at this answer was a little dubious, we can still try plugging it into the equation and seeing if it works.

Example 4.1. Let's consider the separable ODE

$$f'(t) = tf(t)^2.$$

We can write out the formal solution

$$\int \frac{1}{f^2}\,\mathrm{d}f = \int t\,\mathrm{d}t$$

as above. Integrating both sides and remembering to include a constant of integration gives that

$$-\frac{1}{f} = \frac{1}{2}t^2 + C.$$

Letting $c = -C/2$, we can rearrange this to give that

$$f = \frac{2}{c - t^2}.$$

If we want to make sure that this is really a solution to the ODE, we can just differentiate:

$$\left(\frac{2}{c-t^2}\right)' = \frac{4t}{(c-t^2)^2} = tf(t)^2$$

as desired. Thus, for each $c < 0$ we have a maximal solution $(\mathbb{R}, \frac{2}{c-t^2})$, for $c = 0$ we have two maximal solutions

$$\left((-\infty, 0), -\frac{2}{t^2}\right) \qquad \text{and} \qquad \left((0, \infty), -\frac{2}{t^2}\right)$$

while for each $c > 0$ we have three maximal solutions

$$\left((-\infty, -\sqrt{c}), \frac{2}{c-t^2}\right), \qquad \left((-\sqrt{c}, \sqrt{c}), \frac{2}{c-t^2}\right), \quad \text{and} \quad \left((\sqrt{c}, \infty), \frac{2}{c-t^2}\right).$$

These solutions are maximal since the functions blow-up as they approach $\pm\sqrt{c}$ and so cannot be extended to any larger interval. Of course $(\mathbb{R}, 0)$ is also a maximal solution (which can be thought of as the limit of the other solutions when we take $c \to \pm\infty$). Are these the only maximal solutions? Yes! Since the function $F(t, x) = tx^2$ is continuously differentiable it is locally Lipschitz, and Global Picard-Lindeöf implies that there is exactly one maximal solution passing through each space-time point (t_0, x_0), and since every non-zero x_0 can be written as $2/(c-t_0^2)$ for appropriate choice of $c \in \mathbb{R}$ the solutions we have just written down are the only maximal solutions that exist. (In particular, every other solution is just a restriction of one of these solutions to a smaller domain.)

Exercise 42. Find all maximal solutions to the ODE $f' = t(f-1)^2$. (Hint: You don't have to redo all the work we just did!)

Exercise 43. Find all maximal solutions to the ODE $f' = t^3 f^4$.

Let's now return to the general form of our separable ODE $f' = F(t)G(f)$. Since we want to express f in terms of t, we should be careful to restrict to an interval where the antiderivative of $1/G$ is injective. Suppose therefore that F and G are continuous and that I_1 is such that $G \neq 0$ on I_1 and $\Phi : I_1 \to I_2$ is an antiderivative of $1/G$ that is a bijection between its range and its domain, and that $\Psi : I_3 \to \mathbb{R}$ is an antiderivative of F. If $\Psi(x) \in I_2$ for every $x \in I_3$ then we can safely define a function $f : I_3 \to \mathbb{R}$ by

$$f(t) = \Phi^{-1}(\Psi(t)),$$

which is a formal way of solving for f in the equation (4.1). We can easily check that this solves the ODE $f' = F(t)G(f(t))$:

$$f'(t) = (\Phi^{-1})'(\Psi(t))\Psi'(t) = \frac{1}{\Phi'(\Phi^{-1}(\Psi(t)))}\Psi'(t) = G(f(t))F(t)$$

as required, where we used that the derivative of an inverse function satisfies

$$(\Phi^{-1})'(t) = \frac{1}{\Phi'(\Phi^{-1}(t))}.$$

(If you don't remember why this is true you should prove it. Note that if you assume that Φ^{-1} is differentiable then its derivative must be given by this formula by the chain rule since $(\Phi^{-1} \circ \Phi)' = 1$.)

This method always gives us *some* solutions, provided that G is not always zero. Indeed, if G is not always zero then, since it is continuous, there are non-trivial intervals on which G is non-zero with constant sign, and on such intervals the antiderivative of $1/G$ is strictly monotone and hence invertible. Are they the only solutions? Not necessarily! We already saw in our previous example, which was a very nice ODE satisfying the hypotheses of Global Picard-Lindelöf, that we missed the constant zero solution when solving the ODE this way. Let's see what happens in a situation we know to be problematic.

Example 4.2. If we take $F(t) \equiv 1$ and $G(x) \equiv x^{2/3}$ then our separable ODE is just the ODE

$$f' = f^{2/3}.$$

We can write out the formal solution

$$\int \frac{1}{f^{2/3}} \, \mathrm{d}f = \int 1 \, \mathrm{d}t$$

as above. Integrating both sides and remembering to include a constant of integration gives that

$$3f^{1/3} = t + C,$$

and solving for f gives that

$$f(t) = \frac{1}{27}(t + C)^3.$$

While this does give a family of solutions, we are missing both the constant zero solution, the solutions of the form $\frac{1}{27}\max\{0, (t + C)^3\}$ that stick to zero rather than becoming negative, and solutions of the form $\frac{1}{27}\min\{0, (t+C)^3\}$ that stick to zero rather than becoming positive. Yet another kind of solution is given by piecewise functions of the form

$$f(t) = \begin{cases} \frac{1}{27}(t-b)^3 & t > b \\ 0 & a \leq t \leq b \\ \frac{1}{27}(t-a)^3 & t < a. \end{cases}$$

This is all despite the antiderivative of $1/G$ being well-defined and bijection $\mathbb{R} \to \mathbb{R}$, which is as nice as we could hope for for the method otherwise. With care, one can still use the Global Picard-Lindelöf theorem to prove that these are the only solutions:

Exercise 44. Prove that every maximal solution to the ODE $f' = f^{2/3}$ is of one of the forms just listed.

Exercise 45. Let $\phi : \mathbb{R} \to (0, \infty)$ be an increasing, continuously differentiable function. Prove that the initial value problem

$$f(0) = 1 \qquad f' = \phi(f)$$

has a solution defined on all of $\mathbb{R}$ if and only if

$$\int_1^\infty \frac{1}{\phi(t)} \, \mathrm{d}t = \infty.$$

Example 4.3. Let's now consider an example where the antiderivative of $1/G$ is not bijective, namely

$$f' = \frac{1}{f},$$

where $F \equiv 1$ and $G(x) = 1/x$ is defined on $\mathbb{R} \setminus \{0\}$. We can write down the formal solution

$$\int f \, \mathrm{d}f = \int 1 \, \mathrm{d}t$$

which leads to

$$\frac{1}{2} f^2 = t + C$$

for a constant $C \in \mathbb{R}$. Taking the square root leads to two families of maximal solutions

$$\left((-C, \infty), -\sqrt{2t + 2C}\right) \qquad \text{and} \quad \left((-C, \infty), +\sqrt{2t + 2C}\right),$$

so that we get a pair of maximal solutions for each constant $C \in \mathbb{R}$. These solutions really are maximal, since extending them continuously to $-C$ would give a zero value of f where the ODE is not defined. They are also the *only* maximal solutions of the ODE since $1/x$ is a continuously differentiable function of t and x on the phase space $\Omega = \{(t, x) : x \neq 0\}$ and solutions of this form pass through every phase-space point.

5 The Laplace transform

5.1 Definition and basic properties

In this section we study the **Laplace transform**, a tool which sometimes lets us solve ODEs by solving equivalent *algebraic* equations. This method is very powerful, and was the most popular way for engineers to solve ODEs by hand before the advent of widespread computing in the second half of the 20th century. We will see that it can also be very useful for extracting *large time asymptotics* on solutions even when we cannot solve the equation explicitly.

Given a continuous function $f : (0, \infty) \to \mathbb{R}$, the **Laplace transform** $\mathcal{L}\{f\}$ is the function with domain $\{s \in \mathbb{R} : \int_0^\infty |f(t)|e^{-st}\,\mathrm{d}t < \infty\}$ defined by

$$\mathcal{L}\{f\}(s) = \int_0^\infty f(t)e^{-st}\,\mathrm{d}t.$$

Example 5.1. $\mathcal{L}\{1\}$ has domain $(0, \infty)$ and is given by $\mathcal{L}\{1\}(s) = \frac{1}{s}$ for every $s > 0$.

Remark 5.2. It is not at all important that f is continuous, and you could instead take f to be e.g. Riemann integrable when restricted to any closed bounded interval in $(0, \infty)$. In fact one can also unproblematically define the Laplace transform of things that aren't even really functions, such as the Dirac delta function. These generalizations are important in applications, but we will avoid dealing with them since we have not set up all the relevant background. If you learn some measure theory in a subsequent course you will be able to revisit these notes and see that everything works for Laplace transforms of, say, locally finite measures on $(0, \infty)$.

Exercise 46. Prove that $\mathcal{L}\{f\}$ is continuous on its domain of definition.

The domain of $\mathcal{L}\{f\}$ is always either empty (which will never happen in examples we are interested in) or is an interval of the form $(s_*(f), \infty)$ or $[s_*(f), \infty)$ where

$$s_*(f) = \inf\{s \in \mathbb{R} : \int_0^\infty |f(t)|e^{-st}\,\mathrm{d}t < \infty\}.$$

Note in particular that if $a \in \mathbb{R}$ and $C \geq 0$ are such that $|f(t)| \leq Ce^{at}$ for every $t > 0$ then $s_*(f) < a$, so that the domain of $\mathcal{L}\{f\}$ is non-empty. Such a function is said to be of **exponential type**.

The Laplace transform satisfies a large number of useful identities. We will now go through the most useful of them. The first is trivial from linearity of integration.

Lemma 5.3 (Linearity)**.** *If f and g are two continuous functions $f, g : (0, \infty) \to \mathbb{R}$ and s belongs to the domain of both $\mathcal{L}\{f\}$ and $\mathcal{L}\{g\}$ then s belongs to the domain of $\mathcal{L}\{af + bg\}$ and $\mathcal{L}\{af + bg\}(s) = a\mathcal{L}\{f\}(s) + b\mathcal{L}\{g\}(s)$ for every $a, b \in \mathbb{R}$.*

Another obvious identity is as follows.

Lemma 5.4 (Multiplication by exponential $\longrightarrow$ shifting s)**.** *If $f : (0,\infty) \to \mathbb{R}$ is continuous and $a \in \mathbb{R}$ then $\mathcal{L}\{e^{at}f\}$ has domain $\{s + a : s \text{ is in the domain of } \mathcal{L}\{f\}\}$ and satisfies $\mathcal{L}\{e^{at}f\}(s) = \mathcal{L}\{f\}(s-a)$ for every s in the domain of $\mathcal{L}\{e^{at}f\}$.*

The next identity accounts for most the usefulness of Laplace transforms for solving ODEs.

Lemma 5.5 (Differentiation $\longrightarrow$ multiplication by s)**.** *If f is a continuously differentiable function $f : (0,\infty) \to \mathbb{R}$ and $f(0+) = \lim_{t\downarrow 0} f(t)$ is well-defined then*

$$\mathcal{L}\{f'\}(s) = s\mathcal{L}\{f\}(s) - f(0+)$$

for every s in the domain of both $\mathcal{L}\{f'\}$ and $\mathcal{L}\{f\}$.

Remark 5.6. If $f = \sin(e^{t^2})$ then $\mathcal{L}\{f\}$ has domain $(0,\infty)$ but $\mathcal{L}\{f'\}$ has empty domain. Of course this situation can be rescued by weakening our requirement that all integrals converge absolutely in the domain of the Laplace transform. On the other hand it's not worth worrying too much about treating such pathological examples since they do not tend to arise in applications. In most 'nice' examples we will have that $s_*(f') = s_*(f)$.

Proof. This is just integration by parts! If $0 < a < b$ and s belongs to the domain of both $\mathcal{L}\{f\}$ and $\mathcal{L}\{f'\}$ then

$$\int_a^b f'(t)e^{-st}\,\mathrm{d}t = f(b)e^{-sb} - f(a)e^{-sa} - \int_a^b f(t)(e^{-st})'\,\mathrm{d}t = s\int_a^b f(t)e^{-st}\,\mathrm{d}t + f(b)e^{-sb} - f(a)e^{-sa}$$

and the claim follows by taking the limit as $a \downarrow 0$ and $b \uparrow \infty$, which we can do unproblematically since all the relevant integrals converge absolutely. The only thing that requires further justification is that we can get rid of the term $f(b)e^{-sb}$. To do this, we just note that if $\inf_{u\geq t} |f(u)e^{-su}|$ is positive for some t then the integral $\int_t^\infty |f(u)|e^{-su}du$ must be infinite, and since $\int_0^\infty |f(u)|e^{-su}du$ is finite by assumption we must have that $\inf_{u\geq t} |f(u)e^{-su}| = 0$ for every $t > 0$, so that we can find a sequence b_n with $b_n \to \infty$ such that $f(b_n)e^{-sb_n} \to 0$. □

Corollary 5.7 (Repeated differentiation)**.** *If f is an n-times continuously differentiable function $f : (0,\infty) \to \mathbb{R}$ and $f^{(m)}(0+) = \lim_{t\downarrow 0} f^{(m)}(t)$ is well-defined for every $0 \leq m \leq n-1$ then*

$$\mathcal{L}\{f^{(n)}\}(s) = s^n\mathcal{L}\{f\}(s) - \sum_{i=0}^{n-1} s^{n-i-1}f^{(i)}(0+).$$

for every s belonging to the domain of $\mathcal{L}\{f^{(i)}\}$ for every $0 \leq i \leq n$.

Proof. This follows from the differentiation identity by induction on n. Indeed, given that the claim holds for the $(n-1)$th derivative, we deduce that

$$\begin{aligned}\mathcal{L}\{f^n\} &= s\mathcal{L}\{f^{(n-1)}\} - f^{(n-1)}(0+)\\ &= s\left(s^{n-1}\mathcal{L}\{f\} - \sum_{i=0}^{n-2} s^{n-i-2}f^{(i)}(0+)\right) - f^{(n-1)}(0+) = s^n\mathcal{L}\{f\}(s) - \sum_{i=0}^{n-1} s^{n-i-1}f^{(i)}(0+)\end{aligned}$$

for every s in the domain of $\mathcal{L}\{f^{(i)}\}$ for every $0 \leq i \leq n$ as claimed. $\square$

Exercise 47. Prove that if $f : (0, \infty) \to \mathbb{R}$ is a continuous function such that $\int_0^1 |f(t)| \, \mathrm{d}t < \infty$ then

$$\mathcal{L}\left\{\int_0^t f(u) \, \mathrm{d}u\right\}(s) = \frac{1}{s}\mathcal{L}\{f\}(s)$$

for every s in the domain of both Laplace transforms.

Given constants $a_{n-1}, \ldots, a_0$ and $b : (0, \infty) \to \mathbb{R}$ continuous, it follows that any function $f : (0, \infty) \to \mathbb{R}$ solving the constant coefficient linear ODE (with possibly non-constant inhomogeneity b)

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

with $f^{(m)}(0+)$ well-defined for every $0 \leq m \leq n-1$ must satisfy

$$s^n\mathcal{L}\{f\}(s) - \sum_{i=0}^{n-1} s^{n-1-i}f^{(i)}(0+) + a_{n-1}\left(s^{n-1}\mathcal{L}\{f\}(s) - \sum_{i=0}^{n-2} s^{n-2-i}f^{(i)}(0+)\right) + \cdots + a_0\mathcal{L}\{f\}(s) = \mathcal{L}\{b\}(s)$$

for every s in the domain of both $\mathcal{L}\{b\}$ and $\mathcal{L}\{f^{(i)}\}$ for every $0 \leq i \leq n-1$, so that

$$s^n\mathcal{L}\{f\}(s) + a_{n-1}s^{n_1}\mathcal{L}\{f\}(s) + \cdots + a_0\mathcal{L}\{f\}(s) = \mathcal{L}\{b\}(s) + P(s)$$

for some polynomial P of degree at most $n-1$ whose coefficients are determined by the initial conditions $(f^{(i)}(0+) : 0 \leq i \leq n-1)$. Rearranging, it follows that

$$\mathcal{L}\{f\}(s) = \frac{\mathcal{L}\{b\}(s) + P(s)}{s^n + a_{n-1}s^{n-1} + \cdots + a_0}$$

for every s in the domain of both $\mathcal{L}\{b\}$ and $\mathcal{L}\{f^{(i)}\}$ for every $0 \leq i \leq n-1$. This already hints at the power of the Laplace transform since – assuming the Laplace transform is invertible! – we can solve the ODE by solving an *algebraic* equation for $\mathcal{L}\{f\}$ then inverting. Of course this is not so exciting since we already know how to solve constant coefficient linear ODEs, but it does hint at the power of the method.

Of course if we want to make sure this really gives a solution of our ODE, we need to make sure that we can invert the Laplace transform.

Theorem 5.8 (Injectivity of the Laplace transform)**.** *Let $f, g : (0, \infty) \to \mathbb{R}$ be continuous functions of exponential type. If there exists $s_0 \in \mathbb{R}$ such that $\mathcal{L}\{f\}$ and $\mathcal{L}\{g\}$ are defined and equal to each other on (s_0, ∞) then $f = g$.*

We will prove this theorem using the following fact. This fact is an easy consequence of the *Weierstrass approximation theorem*, which states that polynomials are dense in $C([0, 1], \mathbb{R})$.

(That is, any continuous function from $[0,1] \to \mathbb{R}$ can be written as a $\|\cdot\|_\infty$-limit of polynomials.)

Fact 5.9. *Prove that if $h : [0,1] \to \mathbb{R}$ is a continuous function such that $\int_0^1 h(x)P(x)\,\mathrm{d}x = 0$ for every polynomial P, then $h \equiv 0$.*

Proof of Theorem 5.8. Suppose that $f, g : (0,\infty) \to \mathbb{R}$ are two continuous functions of exponential type whose Laplace transforms are defined and equal to each other on some interval (s_0,∞). By increasing if s_0 if necessary, we may assume that $\lim_{t\to\infty} f(t)e^{-s_0 t} = \lim_{t\to\infty} g(t)e^{-s_0 t} = 0$. Setting $u = f - g$, we have that $\mathcal{L}\{u\}$ is defined and equal to zero on (s_0,∞) and that $\lim_{t\to\infty} u(t)e^{-s_0 t} = 0$. We can therefore define a continuous function $h : [0,1] \to \mathbb{R}$ by $h(x) = x^{s_0}u(-\log x)$ for $x \in (0,1]$ and $h(0) = 0$. Moreover, we have by a change of variables $t = -\log x$ (so that $x = e^{-t}$ and $e^{-t}dt = dx$) that

$$0 = \mathcal{L}\{u\}(s_0 + n + 1) = \int_0^\infty u(t)e^{-s_0 t}e^{-nt}e^{-t}\,\mathrm{d}t = \int_0^1 x^n h(x)\,\mathrm{d}x$$

for every $n \geq 0$. It follows from the above exercise that $h \equiv 0$ and hence that $f \equiv g$. □

Unfortunately it is usually very hard (i.e. impossible) to explicitly invert a Laplace transform. The situation is closely analogous to symbolic integration – differentiating symbolically and applying the Laplace transform symbolically are both relatively easy, but to go in the other direction we usually just have to recognize our function as the derivative/Laplace transform of a function we already know. We will see later that the Laplace transform method can also be very useful to extract *large time asymptotics* of solutions even when closed-form solutions are not available.

In order to apply the Laplace transform method, we need to build up a good supply of functions whose Laplace transforms we know. Let's start with the simplest possible thing:

$$\mathcal{L}\{1\} \text{ has domain } (0,\infty) \text{ and is given by } \mathcal{L}\{1\}(s) = \frac{1}{s}.$$

Next, we show that multiplying f by powers of t corresponds to differentiating or integrating $\mathcal{L}\{f\}$.

Lemma 5.10 (Division by t $\longrightarrow$ integration)**.** *Let $f : (0,\infty) \to \mathbb{R}$ be a continuous function such that*

$$\int_0^1 \frac{|f(t)|}{t}\,\mathrm{d}t < \infty.$$

Then the domain of $\mathcal{L}\{f/t\}$ contains that of $\mathcal{L}\{f\}$ and

$$\mathcal{L}\left\{\frac{f}{t}\right\}(s) = \int_s^\infty \mathcal{L}\{f\}(u)\,\mathrm{d}u$$

for every s in the domain of $\mathcal{L}\{f\}$. Moreover, the integral on the right hand side converges absolutely for every such s.

The proof of this identity will use **Fubini's theorem**, which states that if $f : I_1 \times I_2 \to \mathbb{R}$ is a continuous function defined on a product of two non-trivial intervals and $\int_{I_1}\int_{I_2} |f(x,y)|\,\mathrm{d}x\,\mathrm{d}y < \infty$ then

$$\int_{I_1}\int_{I_2} f(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_{I_2}\int_{I_1} f(x,y)\,\mathrm{d}y\,\mathrm{d}x.$$

That is, we can compute a double integral in either order provided it converges absolutely.

Proof of Lemma 5.10. First observe that the hypothesis ensures that

$$\int_0^\infty \frac{|f(t)|}{t} e^{-st}\,\mathrm{d}t \le e^{\max\{-s,0\}} \int_0^1 \frac{|f(t)|}{t}\,\mathrm{d}t + \int_0^\infty |f(t)| e^{-st}\,\mathrm{d}t$$

for every $s \in \mathbb{R}$ and hence that the domain of $\mathcal{L}\{f/t\}$ contains that of $\mathcal{L}\{f\}$ as claimed. Since all the relevant integrals converge absolutely, we can use Fubini to compute that

$$\begin{aligned}\mathcal{L}\{f/t\}(s) = \int_0^\infty \frac{f(t)}{t} e^{-st}\,\mathrm{d}t = \int_0^\infty f(t) \int_s^\infty e^{-ut}\,\mathrm{d}u\,\mathrm{d}t \\ = \int_s^\infty \int_0^\infty f(t) e^{-ut}\,\mathrm{d}t\,\mathrm{d}u = \int_s^\infty \mathcal{L}\{f\}(u)\,\mathrm{d}u\end{aligned}$$

for every s in the domain of $\mathcal{L}\{f\}$ as claimed. □

Lemma 5.11 (Multiplication by t $\longrightarrow$ differentiation)**.** *Let $f : (0,\infty) \to \mathbb{R}$ be a continuous function. Then $s_*(tf) \le s_*(f)$ and*

$$\mathcal{L}\{tf\}(s) = -\frac{d}{ds}\mathcal{L}\{f\}(s)$$

for every $s > s_*(f)$. In particular, $\mathcal{L}\{f\}$ is differentiable on $(s_*(f),\infty)$.

Proof. If $\int_0^1 |f(t)|\,\mathrm{d}t = \infty$ then $\mathcal{L}\{f\}$ has empty domain and the claim is trivial, so we may suppose that this integral is finite. Since $\int_0^1 |f(t)|\,\mathrm{d}t < \infty$, we can apply Lemma 5.10 to tf to deduce that, in this case, the domain of $\mathcal{L}\{f\}$ contains that of $\mathcal{L}\{tf\}$ and that

$$\mathcal{L}\{f\}(s) = \int_s^\infty \mathcal{L}\{tf\}(u)\,\mathrm{d}u$$

for every s in the domain of $\mathcal{L}\{f\}$. The claim follows from the fundamental theorem of calculus. □

Corollary 5.12. *Let $f : (0,\infty) \to \mathbb{R}$ be a continuous function. Then for each $n \ge 1$ we have that $s_*(t^n f) \le s_*(f)$ and*

$$\mathcal{L}\{t^n f\}(s) = (-1)^n \frac{d^n}{ds^n}\mathcal{L}\{f\}(s)$$

for every $s > s_*(f)$. In particular, $\mathcal{L}\{f\}$ is smooth on $(s_*(f),\infty)$.

It follows from this corollary that the Laplace transform $\mathcal{L}\{t^n\}$, which has domain $(0,\infty)$, is given by

$$\mathcal{L}\{t^n\}(s) = (-1)^n \frac{d^n}{ds^n} \int_0^\infty e^{-ts}\,\mathrm{d}t = \frac{n!}{s^{n+1}},$$

for each $n \geq 0$, where we stress that we are thinking of t^n as a function defined on $(0,\infty)$. Together with the fact that multiplication by an exponential corresponds to shifting s, it follows that $\mathcal{L}\{e^{at}t^n\}$ has domain (a,∞) and that

$$\mathcal{L}\{e^{at}t^n\}(s) = \frac{n!}{(s-a)^{n+1}}$$

for each $n \geq 0$ and $a \in \mathbb{R}$. We can also easily compute the Laplace transforms of trig functions.

Lemma 5.13. *For each $\omega \in \mathbb{R}$, $\mathcal{L}\{\sin(\omega t)\}$ and $\mathcal{L}\{\cos(\omega t)\}$ have domain $(0,\infty)$ and are given by*

$$\mathcal{L}\{\sin(\omega t)\}(s) = \frac{\omega}{s^2+\omega^2} \qquad \text{and} \qquad \mathcal{L}\{\cos(\omega t)\}(s) = \frac{s}{s^2+\omega^2}$$

for every $s > 0$.

Proof. We have that

$$\int_0^\infty \cos(\omega t)e^{-st}\,\mathrm{d}t = \Re\int_0^\infty e^{\omega it - st}\,\mathrm{d}t = \Re\frac{1}{s-\omega i} = \Re\frac{s+\omega i}{(s-\omega i)(s+\omega i)} = \Re\frac{s+\omega i}{s^2+\omega^2} = \frac{s}{s^2+\omega^2}$$

as claimed, where the symbol $\Re$ means "the real part of". The computation for sin is similar except we take imaginary parts instead of real parts. □

It follows from these rules that $\mathcal{L}\{e^{at}\sin(\omega t)\}$ and $\mathcal{L}\{e^{at}\cos(\omega t)\}$ have domain (a,∞) and that

$$\mathcal{L}\{e^{at}\sin(\omega t)\}(s) = \frac{\omega}{(s-a)^2+\omega^2} \qquad \text{and} \qquad \mathcal{L}\{e^{at}\cos(\omega t)\}(s) = \frac{s-a}{(s-a)^2+\omega^2}$$

for every $s > a$.

Example 5.14. Let's see how we can use the Laplace transform to solve the damped spring equation $f'' + 2\zeta f' + f = 0$ on the interval $(0,\infty)$. Any solution to the equation that extends to a continuously differentiable function on $[0,\infty)$ must have

$$\mathcal{L}\{f'' + 2\zeta f' + f\} = s^2\mathcal{L}\{f\}(s) - sf(0+) - f'(0+) + 2\zeta s\mathcal{L}\{f\} - 2\zeta f(0+) + \mathcal{L}\{f\} = 0,$$

for every $s > s_*(f)$, which we can rearrange to give that

$$\mathcal{L}\{f\}(s) = \frac{f'(0+) + (s+2\zeta)f(0+)}{s^2+2\zeta s+1}.$$

To proceed, we need to be able to recognise this as the Laplace transform of something we know. This can be done using **partial fractions**.

Recall that a **rational function** is a function of the form $P(x)/Q(x)$ where P and Q are both polynomials. A real polynomial is said to be **irreducible** if it cannot be written as the product of two non-constant real polynomials. The **fundamental theorem of algebra** states that every degree n polynomial can be written uniquely in the form

$$Q(x) = A\prod_{i=1}^{n}(x-\lambda_i)$$

where $A \in \mathbb{C}$ and $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$. If Q is real then the roots λ_i must either be real or come in a complex conjugate pair, and it follows that every real polynomial can be written uniquely as a product of real irreducible polynomials

$$Q(x) = A\prod_{i=1}^{k}(x-\lambda_i)\prod_{j=1}^{\ell}(x^2+\sigma_i x+\omega_i)$$

where $\lambda_1, \ldots, \lambda_k$ are real and $\sigma_1, \ldots, \sigma_\ell$ and $\omega_1, \ldots, \omega_\ell$ are real numbers such that $\sigma_j^2 - 4\omega_j < 0$ for each $1 \leq j \leq \ell$ (of course k and ℓ could be zero). For our purposes, it will be more helpful to group identical terms and write

$$Q(x) = A\prod_{i=1}^{k}(x-\lambda_i)^{n_i}\prod_{j=1}^{\ell}(x^2+\sigma_j x+\omega_j)^{m_j}$$

where the λ_is and the pairs (σ_j, ω_j) are all distinct and n_i, m_j are positive integers. The partial fraction expansion states that we can always write a rational function in the form

$$\frac{P(x)}{Q(x)} = A(x) + \sum_{i=1}^{r}\frac{A_i(x)}{Q_i(x)}$$

where

1. If Q has larger degree than P then $A = 0$. If P and Q have the same degree then A is a constant. If P has larger degree than Q then A is a polynomial of degree $\deg P - \deg Q$.

2. $Q_1, \ldots, Q_r$ are polynomials of the form $(x-\lambda)^n$ or $(x^2+\sigma x+\omega)^n$ that divide Q and are a power of an irreducible polynomial, and

3. $A_1, \ldots, A_r$ are polynomials such that A_i has degree strictly smaller than the irreducible polynomial of which Q_i is a power of for each $1 \leq i \leq k$.

To use partial fraction expansions to solve ODEs one must compute what the polynomials A_i are, and doing this amounts to a linear algebra problem. Of course, we are secretly doing something very similar to computing the Jordan normal form of a matrix, since the two techniques can both be used to compute the general solution to constant coefficient linear ODEs.

Let us illustrate how this works in our simple example. As when we solved the ODE using matrix exponentiation, something different will happen according to whether $s^2+2\zeta s+1$ has two real roots, two complex conjugate roots, or a single real root. In the first case, which occurs when $|\zeta|>1$, we want to write

$$\frac{f'(0+)+(s+2\zeta)f(0+)}{s^2+2\zeta+1}=\frac{a}{s-\zeta-\sqrt{\zeta^2-1}}+\frac{b}{s-\zeta+\sqrt{\zeta^2-1}}$$

where a,b are real. Adding these together we get that

$$\frac{f'(0+)+(s+2\zeta)f(0+)}{s^2+2\zeta+1}=\frac{(s-\zeta)(a+b)+\sqrt{\zeta^2-1}(a-b)}{(s-\zeta-\sqrt{\zeta^2-1})(s-\zeta+\sqrt{\zeta^2-1})},$$

so that, comparing the s terms and constant terms in the numerator,

$$a+b=f(0+)\qquad\text{and}\qquad(\zeta-\sqrt{\zeta^2-1})a+(\zeta+\sqrt{\zeta^2-1})b=-2\zeta f(0+)-f'(0+).$$

This system of linear equations can of course be solved by inverting a 2×2 matrix. This will give us some explicit constants a and b determined by ζ, $f(0+)$, and $f'(0+)$ such that

$$\mathcal{L}\{f\}=\frac{a}{s-\zeta-\sqrt{\zeta^2-1}}+\frac{b}{s-\zeta+\sqrt{\zeta^2-1}}=a\mathcal{L}\{e^{-(\zeta+\sqrt{\zeta^2-1})t}\}+b\mathcal{L}\{e^{-(\zeta-\sqrt{\zeta^2-1})t}\}.$$

Thus, we can deduce that, **assuming that s was in the domain of all relevant Laplace transforms whenever we needed it to be**, our solution must be of the form

$$f(t)=ae^{-(\zeta+\sqrt{\zeta^2-1})t}+be^{-(\zeta-\sqrt{\zeta^2-1})t}.$$

While it is possible to go through and justify this assumption at each step, we can instead just check that this really is a solution to our ODE, and deduce from Picard-Lindelöf that we have got every solution.

If $|\zeta|<1$ then we complete the square to write $s^2+2\zeta s+1=(s+\zeta)^2+1-\zeta^2$

$$\begin{aligned}\mathcal{L}\{f\}&=\frac{f'(0+)+(s+2\zeta)f(0+)}{(s+\zeta)^2+1-\zeta^2}=\frac{f'(0+)+\zeta f(0+)}{(s+\zeta)^2+1-\zeta^2}+\frac{(s+\zeta)f(0+)}{(s+\zeta)^2+1-\zeta^2}\\&=\frac{f'(0+)+\zeta f(0+)}{\sqrt{1-\zeta^2}}\mathcal{L}\{e^{-\zeta t}\sin(t\sqrt{1-\zeta^2})\}+f(0+)\mathcal{L}\{e^{-\zeta t}\cos(t\sqrt{1-\zeta^2})\}.\end{aligned}$$

Again, **assuming s is the domain of all relevant Laplace transforms**, we deduce that our solution must be of the form

$$f(t)=\frac{f'(0+)+\zeta f(0+)}{\sqrt{1-\zeta^2}}e^{-\zeta t}\sin(t\sqrt{1-\zeta^2})+f(0+)e^{-\zeta t}\cos(t\sqrt{1-\zeta^2}).$$

As before, rather than justifying that s was indeed always in the relevant domain, we can just check that this is indeed a solution and deduce from Picard-Lindelöf that every solution is of this form.

Finally, if $|\zeta| = 1$ then $s^2 + 2\zeta s + 1 = (s+\zeta)^2$ and we seek a partial fractions expansion of the form

$$\mathcal{L}\{f\} = \frac{f'(0+) + (s+2\zeta)f(0+)}{(s+\zeta)^2} = \frac{a}{s+\zeta} + \frac{b}{(s+\zeta)^2}.$$

To solve for a and b we can add the two fractions and compare numerators to obtain that

$$as + a\zeta + b = sf(0+) + f'(0+) + 2\zeta f(0+),$$

so that $a = f(0+)$ and $b = f'(0+) + \zeta f(0+)$. Thus, we have that

$$\begin{aligned}\mathcal{L}\{f\} &= \frac{f(0+)}{s+\zeta} + \frac{f'(0+) + \zeta f(0+)}{(s+\zeta)^2} = f(0+)\mathcal{L}\{e^{-\zeta t}\} - (f'(0+) + \zeta f(0+))\mathcal{L}\{e^{-\zeta t}\}' \\ &= f(0+)\mathcal{L}\{e^{-\zeta t}\} + (f'(0+) + \zeta f(0+))\mathcal{L}\{te^{-\zeta t}\}.\end{aligned}$$

As before, **assuming that s belongs to the domain of all relevant Laplace transforms**, we deduce that

$$f(t) = f(0+)e^{-\zeta t} + (f'(0+) + \zeta f(0+))te^{-\zeta t}.$$

Checking that this is indeed a solution, we deduce by Picard-Lindelöf that every solution is of this form.

It has come time to humble ourselves by trying to use the Laplace transform solve a simple linear ODE with non-constant coefficients.

Example 5.15. Consider the linear ODE

$$f'' - tf' - f = 0.$$

Suppose that $f : (0,\infty) \to \mathbb{R}$ is a solution to this ODE with $f(0+)$ and $f'(0+)$ well-defined and let I be the intersection of the domains of $\mathcal{L}\{f\}$, $\mathcal{L}\{f'\}$, $\mathcal{L}\{f''\}$ and $\mathcal{L}\{tf'\}$. (Of course this might be empty, and we will need to come back to this issue later.) If $s \in I$ then we have that

$$\mathcal{L}\{f'' - tf' - f\} = s^2\mathcal{L}\{f\} - f'(0+) - sf(0+) + \frac{d}{ds}\left(s\mathcal{L}\{f\} - f(0+)\right) - \mathcal{L}\{f\} = 0$$

and hence, rearranging, that

$$\mathcal{L}\{f\}' + s\mathcal{L}\{f\} = f(0+) + \frac{1}{s}f'(0+)$$

for every $s \in I$ with $s > 0$. Since this is now a *first-order* linear ODE, we can let $s_0 \in I$ and

write down the solution

$$\begin{aligned}\mathcal{L}\{f\} &= \exp\left[-\int_{s_0}^{s} u\,\mathrm{d}u\right]\left(\mathcal{L}\{f\}(s_0) + \int_{s_0}^{s} \exp\left[\int_0^u v\,\mathrm{d}v\right](f(0+) + \frac{1}{u}f'(0+))\,\mathrm{d}u\right) \\ &= e^{-\frac{1}{2}(s^2-s_0^2)}\left(\mathcal{L}\{f\}(s_0) + \int_{s_0}^{s} e^{\frac{1}{2}(u^2-s_0^2)}(f(0+) + \frac{1}{u}f'(0+))\,\mathrm{d}u\right)\end{aligned}$$

for every $s > 0$ in I. A problem appears: The right hand side doesn't look like the Laplace transform of anything we're familiar[12] with! Despite this, it seems clear that we have achieved *something*, and that the same approach would let us compute the Laplace transform of the solution to any linear ODE where all the coefficients are linear in t. Similarly, we can compute the Laplace transform of an ODE whose coefficients are degree m polynomials in t by solving an mth order ODE whose coefficients are rational functions in t, which may or may not be simpler than what we started with.

Remark 5.16. Analyzing this example via other methods, one can prove that there space of solutions to the linear ODE $f'' - tf' - f = 0$ is spanned by two functions f_1 and f_2 where $\mathcal{L}\{f_1\}$ has domain $(0,\infty)$ and $\mathcal{L}\{f_2\}$ has empty domain! Explicitly, these functions are

$$f_1 = e^{t^2/2}\operatorname{erfc}(t/\sqrt{2}) \qquad \text{and} \qquad f_2 = e^{t^2/2}$$

where

$$\operatorname{erfc}(t) := \frac{2}{\sqrt{\pi}}\int_t^\infty e^{-t^2}\,\mathrm{d}t$$

is the **complementary error function**. As such, our Laplace transform analysis does not, in fact, apply to all solutions of the ODE. In applications this is not necessarily a problem. For example, if one is only interested in solutions that converge to zero as $t \to \infty$ then the Laplace transform of such a solution always has domain containing $(0,\infty)$. Moreover, for second-order linear ODEs the **Wronskian method** allows us to compute a second linearly independent solution to the ODE given a single solution, so that a complete set of solutions to this ODE can be found by a combination of the Laplace transform and Wronskian methods.

Exercise 48 (Stokes equation)**.** Let $f : (0,\infty) \to \mathbb{R}$ be a solution to the ODE $f'' = tf$ and suppose that $f(0+)$ and $f'(0+)$ are well-defined and that the domain of the Laplace transform of f is non-empty. Compute the Laplace transform of f.

Later, we will return to the following question: What can we learn about a function from its Laplace transform, even when we cannot compute the function explicitly?

[12] In fact this ODE does admit a solution by quadrature that can be found by explicitly inverting this Laplace transform, but let that not distract us from the main point that we may wish to extract information from the Laplace transform in situations where we cannot explicitly invert it!

5.2 Convolutions and inhomogeneous linear ODEs

Given constants $a_{n-1}, \dots, a_0$ and $b : (0, \infty) \to \mathbb{R}$ continuous, we saw in the previous section that any function $f : (0, \infty) \to \mathbb{R}$ solving the linear ODE

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

with $f^{(i)}(0+)$ well-defined for every $0 \leq i \leq n-1$ must satisfy

$$\mathcal{L}\{f\}(s) = \frac{\mathcal{L}\{b\}(s) + P(s)}{s^n + a_{n-1}s^{n-1} + \cdots + a_0}$$

for every s in the domain of both $\mathcal{L}\{b\}$ and $\mathcal{L}\{f^{(i)}\}$ for every $0 \leq i \leq n-1$ for which $s^n + a_{n-1}s^{n-1} + \cdots + a_0 \neq 0$, where P is a polynomial of degree at most $n-1$ whose coefficients are determined by the initial conditions $(f^{(i)}(0+) : 0 \leq i \leq n-1)$. This means in particular that we should be able to find a particular solution to the ODE by inverting the Laplace transform

$$\frac{\mathcal{L}\{b\}(s)}{s^n + a_{n-1}s^{n-1} + \cdots + a_0}.$$

We already know how to find the inverse Laplace transform of $(s^n + a_{n-1}s^{n-1} + \cdots + a_0)^{-1}$ using partial fractions. It turns out there is an easy way to invert the product of this with $\mathcal{L}\{b\}$ (or at least to write the inverse as an integral).

Suppose that $f, g : (0, \infty) \to \mathbb{R}$ are continuous with $\int_0^1 |f(t)|\,\mathrm{d}t, \int_0^1 |g(t)|\,\mathrm{d}t < \infty$. The **convolution** of f and g, denoted $f * g$, is the function $f * g : (0, \infty) \to \mathbb{R}$ defined by

$$f * g(t) = \int_0^t f(\tau) g(t - \tau)\,\mathrm{d}\tau,$$

which is well-defined since

$$\begin{aligned}\int_0^t |f(\tau)||g(t-\tau)|\,\mathrm{d}\tau &= \int_0^{t/2} |f(\tau)||g(t-\tau)|\,\mathrm{d}\tau + \int_{t/2}^t |f(\tau)||g(t-\tau)|\,\mathrm{d}\tau \\ &\leq \sup_{t/2 \leq u \leq t} |g(u)| \int_0^{t/2} |f(\tau)|\,\mathrm{d}\tau + \sup_{t/2 \leq u \leq t} |f(u)| \int_0^{t/2} |g(\tau)|\,\mathrm{d}\tau < \infty\end{aligned}$$

for every $t > 0$.

Example 5.17. If $f : (0, \infty) \to \mathbb{R}$ is such that $\int_0^1 |f(t)|\,\mathrm{d}t < \infty$ then $(f * 1)(t) = \int_0^t f(t)\,\mathrm{d}t$.

Example 5.18. We can use the binomial theorem to compute the convolution of two powers

of t:

$$(t^n * t^m) = \int_0^t \tau^n (t-\tau)^m = \sum_{k=0}^{m} (-1)^k \binom{m}{k} t^{m-k} \int_0^t \tau^{n+k} \, \mathrm{d}\tau$$
$$= \left(\sum_{k=0}^{m} \binom{m}{k} \frac{(-1)^k}{n+k+1} \right) t^{n+m+1}.$$

We will shortly perform this calculation another way, using the Laplace transform, and obtain a simpler expression for this constant.

Exercise 49. Suppose that $f, g, h : (0, \infty) \to \mathbb{R}$ are continuous with $\int_0^1 |f(t)| \, \mathrm{d}t$, $\int_0^1 |g(t)| \, \mathrm{d}t$, $\int_0^1 |h(t)| \, \mathrm{d}t < \infty$. Prove the following properties of convolution:

1. Commutativity: $f * g = g * f$.
2. Distributivity: $f * (g + h) = f * g + f * h$.
3. Associativity: $\int_0^1 |(g * h)(t)| \, \mathrm{d}t < \infty$ and $f * (g * h) = (f * g) * h$. (Hint: use Fubini's theorem.)
4. Product rule: If f is continuously differentiable with $\int_0^1 |f'(t)| \, \mathrm{d}t$ and $g(0+)$ is well-defined then $f * g$ is differentiable with $(f * g)' = g(0+)f + f' * g$.

Lemma 5.19 (Factoring out exponentials)**.** *Suppose that $f, g, h : (0, \infty) \to \mathbb{R}$ are continuous with $\int_0^1 |f(t)| \, \mathrm{d}t$ and $\int_0^1 |g(t)| \, \mathrm{d}t$ finite. Then*

$$(e^{-st} f) * (e^{-st} g) = e^{-st} (f * g)$$

for every $s \in \mathbb{R}$.

Proof. We can compute that

$$(e^{-st} f) * (e^{-st} g)(t) = \int_0^t e^{-s\tau} f(\tau) e^{-s(t-\tau)} g(t-\tau) \, \mathrm{d}\tau = e^{-st} \int_0^t f(\tau) g(t-\tau) \, \mathrm{d}\tau = e^{-st} (f * g)$$

as claimed. □

Lemma 5.20 (Integrals of convolutions)**.** *If $f, g : (0, \infty) \to \mathbb{R}$ are continuous with $\int_0^\infty |f(t)| \, \mathrm{d}t$ and $\int_0^\infty |g(t)| \, \mathrm{d}t$ finite then*

$$\int_0^\infty (f * g)(t) \, \mathrm{d}t = \left(\int_0^\infty f(t) \, \mathrm{d}t \right) \cdot \left(\int_0^\infty g(t) \, \mathrm{d}t \right).$$

Proof. Writing $\mathbb{1}(\tau \leq t)$ for the function that is 1 when $\tau \leq t$ otherwise, and using that Fubini remains valid for piecewise-continuous functions, we can compute that

$$\begin{aligned}\int_0^\infty (f*g)(t)\,\mathrm{d}t &= \int_0^\infty \int_0^t f(\tau)g(t-\tau)\,\mathrm{d}\tau\,\mathrm{d}t \\ &= \int_0^\infty \int_0^\infty f(\tau)g(t-\tau)\mathbb{1}(\tau \leq t)\,\mathrm{d}\tau\,\mathrm{d}t \\ &= \int_0^\infty \int_{-\tau}^\infty f(\tau)g(u)\mathbb{1}(0 \leq u)\,\mathrm{d}u\,\mathrm{d}\tau \\ &= \int_0^\infty \int_0^\infty f(\tau)g(u)\,\mathrm{d}u\,\mathrm{d}\tau \\ &= \left(\int_0^\infty f(t)\,\mathrm{d}t\right)\cdot\left(\int_0^\infty g(t)\,\mathrm{d}t\right).\end{aligned}$$

as claimed, where to verify that our application of Fubini was legitimate we do essentially the same calculation to check that

$$\begin{aligned}\int_0^\infty \int_0^\infty |f(\tau)g(t-\tau)|\mathbb{1}(\tau \leq t)\,\mathrm{d}\tau\,\mathrm{d}t &= \int_0^\infty \int_{-\tau}^\infty |f(\tau)g(u)|\mathbb{1}(0 \leq u)\,\mathrm{d}u\,\mathrm{d}\tau \\ &= \int_0^\infty \int_0^\infty |f(\tau)g(u)|\,\mathrm{d}u\,\mathrm{d}\tau \\ &= \left(\int_0^\infty |f(t)|\,\mathrm{d}t\right)\cdot\left(\int_0^\infty |g(t)|\,\mathrm{d}t\right) < \infty. \qquad \square\end{aligned}$$

Theorem 5.21 (Convolution $\longrightarrow$ products)**.** *Suppose that $f, g : (0,\infty) \to \mathbb{R}$ are continuous with $\int_0^1 |f(t)|\,\mathrm{d}t, \int_0^1 |g(t)|\,\mathrm{d}t < \infty$. If s belongs to the domain of both $\mathcal{L}\{f\}$ and $\mathcal{L}\{g\}$ then it belongs to the domain of $\mathcal{L}\{f*g\}$ and*

$$\mathcal{L}\{f*g\}(s) = \mathcal{L}\{f\}(s)\mathcal{L}\{g\}(s).$$

Proof. We first check that s belongs to the domain of $\mathcal{L}\{f*g\}$ whenever it belongs to the domain of $\mathcal{L}\{f\}$ and $\mathcal{L}\{g\}$. Since s belongs to the domain of $\mathcal{L}\{f\}$ and $\mathcal{L}\{g\}$ we have that

$$\begin{aligned}\int_0^\infty |(f*g)(t)|e^{-st}\,\mathrm{d}t &\leq \int_0^\infty (|f|*|g|)(t)e^{-st}\,\mathrm{d}t \\ &= \int_0^\infty ((e^{-st}|f|)*(e^{-st}|g|))(t)\,\mathrm{d}t \\ &= \int_0^\infty e^{-st}|f(t)|\,\mathrm{d}t \int_0^\infty e^{-st}|g(t)|\,\mathrm{d}t < \infty\end{aligned}$$

so that s belongs to the domain of $\mathcal{L}\{f*g\}$ as claimed. Since all relevant integrals converge

absolutely, we can compute that

$$\mathcal{L}\{f*g\}(s) = \int_0^\infty (f*g)(t)e^{-st}\,\mathrm{d}t = \int_0^\infty ((e^{-st}f)*(e^{-st}g))(t)\,\mathrm{d}t$$
$$= \int_0^\infty e^{-st}f(t)\,\mathrm{d}t \int_0^\infty e^{-st}g(t)\,\mathrm{d}t = \mathcal{L}\{f\}(s)\mathcal{L}\{g\}(s)$$

as claimed. □

This means that if s belongs to the domain of all relevant Laplace transforms and

$$\mathcal{L}\{f\} = \frac{\mathcal{L}\{b\}(s) + P(s)}{s^n + a_{n-1}s^{n-1} + \cdots + a_0}$$

then

$$f = b * \mathcal{L}^{-1}\left\{\frac{1}{s^n + a_{n-1}s^{n-1} + \cdots + a_0}\right\} + \mathcal{L}^{-1}\left\{\frac{P(s)}{s^n + a_{n-1}s^{n-1} + \cdots + a_0}\right\}$$

where we have seen how to compute the inverse Laplace transforms of these rational functions using partial fractions (or by solving the relevant homogeneous linear ODE using matrix exponentiation). In other words, the solution to a linear ODE with constant $a_{n-1}, \ldots, a_0$ but possibly non-constant b can be expressed as the convolution of b with the solution the associated homogeneous equation.

Exercise 50. Explain why this solution is the same as that given by Duhamel's principle.

Exercise 51. Use this method to solve the inhomogeneous linear ODE $f' - f = \sin(t)$.

Example 5.22. We have that

$$\mathcal{L}\{t^n * t^m\}(s) = \mathcal{L}\{t^n\}\mathcal{L}\{t^m\} = \frac{n!m!}{s^{n+m+2}} = \frac{n!m!}{(n+m+1)!}\mathcal{L}\{t^{n+m+1}\}$$

and hence that

$$t^n * t^m = \frac{n!m!}{(n+m+1)!}t^{n+m+1}.$$

Comparing this with our direct calculation of the convolution yields the non-obvious (to me!) combinatorial identity

$$\frac{n!m!}{(n+m+1)!} = \sum_{k=0}^{m}\binom{m}{k}\frac{(-1)^k}{n+k+1}.$$

Using Laplace transforms in this way turns out to be a very useful way of proving identities like this, something we will return to in the next section.

Remark 5.23. If X and Y are independent $(0,\infty)$-valued random variables with probability density functions f_X and f_Y then their sum $X+Y$ has probability density function $f_{X+Y} = f_X * f_Y$.

Remark 5.24. One thing Laplace transforms are very useful for is solving functional equations involving both derivatives and convolutions, such as $f' = f * f$. Indeed, if s belongs to the domain of the Laplace transform for some f with $f' = f * f$ then

$$s\mathcal{L}\{f\} - f(0+) = \mathcal{L}\{f\}^2$$

and we can solve the quadratic

$$\mathcal{L}\{f\}(s) = \frac{s \pm \sqrt{s^2 + 4f(0+)}}{2}.$$

Of course to give an explicit solution one would have to find the inverse Laplace transform of this function. At this point it is probably not apparent why one would ever encounter such an equation $f' = f * f$ in the first place, but we will see in the next section that expressions of this form often arise when using ODE techniques in counting problems.

6 Series solutions

In this section we begin to develop the theory of *series solutions*, one of the most powerful and general methods for solving ODEs. We will begin by discussing standard power series solutions of the form $\sum_{n=0}^{\infty} a_n x^n$, but we will later see that one often wants to consider other kinds of series solutions also.

6.1 Formal power series

A formal power series is a series $\sum_{n=0} a_n x^n$ that is considered *as an algebraic object only*, without any consideration of whether the series actually converges and defines a function. The space of (real) formal power series is written $\mathbb{R}[[x]]$. As a *set*, it is in bijection with the set of sequences $\{(a_0, a_1, \ldots) : a_0, a_1, \ldots \in \mathbb{R}\}$; two formal power series are considered to be equal if and only if all their coefficients are equal. The space is a vector space with addition and scalar multiplication defined by

$$\lambda \sum_{n=0}^{\infty} a_n x^n + \mu \sum_{n=0}^{\infty} b_n x^n = \sum_{n=0}^{\infty} (\lambda a_n + \mu b_n) x^n$$

for every two formal power series $\sum_{n=0}^{\infty} a_n x^n$ and $\sum_{n=0}^{\infty} b_n x^n$ and real numbers $\lambda, \mu \in \mathbb{R}$. We can also define multiplication of formal power series by

$$\left(\sum_{n=0}^{\infty} a_n x^n \right) \left(\sum_{n=0}^{\infty} b_n x^n \right) = \sum_{n=0}^{\infty} \left(\sum_{k=0}^{n} a_k b_{n-k} \right) x^n.$$

We stress that this is the **definition** of multiplication for formal power series, and makes sense even if none of the relevant series converge. If you find this confusing, it might help you to think that we are really defining an operation on sequences that takes the pair of sequences (a_n) and (b_n) to the sequence $(\sum_{k=0}^{n} a_k b_{n-k})$.

The formula for the product of two formal power series is closely related to the convolution identity for Laplace transforms: If we define the convolution $a * b$ of two sequences $a = (a_n)$ and $b = b_n$ by $(a * b)_n = \sum_{k=0}^{n} a_k b_{n-k}$ then we can rewrite our multiplication rule as

$$\left(\sum_{n=0}^{\infty} a_n x^n \right) \left(\sum_{n=0}^{\infty} b_n x^n \right) = \sum_{n=0}^{\infty} (a * b)_n x^n.$$

The following exercise gives some legitimacy to the idea that the two notions of convolution we have introduced are discrete and continuous analogues of each other.

Exercise 52 (Discrete and continuous convolutions)**.** Let $f, g : (0, \infty) \to \mathbb{R}$ be continuous

functions extending continuously to 0. Prove that

$$f * g(t) = \lim_{m\to\infty} \frac{1}{m} \left(((f(n/m))_{n\geq 0}) * ((g(n/m))_{n\geq 0})\right)_{\lceil mt \rceil}$$

for every $t > 0$.

Note that polynomials can be thought of as formal power series for which $a_n = 0$ for all sufficiently large n, and that the rules we have defined for addition and multiplication coincide with the usual addition and multiplication rules for polynomials. In particular, we have *by definition* that if $\sum_{n=0}^{\infty} a_n x^n$ is a formal power series then

$$x^m \sum_{n=0}^{\infty} a_n x^n = \left(\sum_{n=0}^{\infty} \mathbb{1}(n = m) x^n\right) \left(\sum_{n=0}^{\infty} a_n x^n\right) = \sum_{n=0}^{\infty} \mathbb{1}(n \geq m) a_{n-m} x^n,$$

where $\mathbb{1}(n \geq m)$ is 1 if $n \geq m$ and 0 if $n < m$.

We say that a formal power series $\sum_{n=0} b_n x^n$ is the **reciprocal** of a formal power series $\sum_{n=0}^{\infty} a_n x^n$ if $\left(\sum_{n=0}^{\infty} a_n x^n\right) \left(\sum_{n=0}^{\infty} b_n x^n\right) = 1$ as formal power series, i.e., if $\sum_{k=0}^{n} a_k b_{n-k}$ is equal to 1 for $n = 0$ and 0 for every $n > 0$.

Proposition 6.1. *A formal power series $\sum_{n=0} a_n x^n$ has a reciprocal if and only if $a_0 \neq 0$, and in this case its reciprocal is unique.*

Proof. Since $\sum_{k=0}^{n} a_k b_{n-k} = a_0 b_0$ when $n = 0$, it is clear that a formal power series with $a_0 = 0$ cannot have a reciprocal, and for any other formal power series, every reciprocal must have $b_0 = a_0^{-1}$. For a formal power series with $a_0 \neq 0$, the formal power series $\sum_{n=0}^{\infty} b_n x^n$ is a reciprocal if and only if $b_0 = a_0^{-1}$ and

$$b_n = -\frac{1}{a_0} \sum_{k=1}^{n} a_k b_{n-k}$$

for every $n \geq 1$, so that each b_n is uniquely determined by $a_0, \ldots, a_n$ and $b_0, \ldots, b_{n-1}$. □

There are many more operations we can define purely algebraically for formal power series. For us the most important will be differentiation:

$$\frac{d}{dx} \left(\sum_{n=0}^{\infty} a_n x^n\right) = \sum_{n=0}^{\infty} (n+1) a_{n+1} x^n.$$

This coincides with usual differentiation whenever we are inside the radius of convergence of the relevant power series (since in this case we can safely differentiate term-by-term as we saw earlier), but also makes sense as a purely algebraic operation on formal power series. Again, if you find the notation confusing, you can think of this as an operation on sequences defined by

$$(a_0, a_1, a_2, \ldots) \mapsto (a_1, 2a_2, 3a_3, \ldots).$$

One can also define integration of formal power series similarly.

It will often be useful to use function-style notation for discussing power series, even if they are not really functions, so that we can write e.g. $f(x) = \sum_{n=0}^{\infty} a_n x^n$ for a formal power series $f(x) \in \mathbb{R}[[x]]$. Of course when you do this you should make sure that you are not doing any "illegal" operations that are not defined formally.

Formal power series from smooth functions. Given a smooth function f defined on an open interval containing 0, we can always consider the Taylor series of f as a formal power series

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

Be warned however that 1) this power series might not converge for any $x \neq 0$, in which case it does not define a function and 2) even it it does converge, the function it defines might not be equal to f. We will return to examples of the first kind later in the course. For a simple example of the second kind, one can consider the function

$$f(x) = \begin{cases} e^{-x^{-2}} & x \neq 0 \\ 0 & x = 0. \end{cases}$$

This function is smooth and has $f^{(n)}(0) = 0$ for every $n \geq 0$, but is not identically equal to zero in any neighbourhood of zero. This particular function is not special: any smooth function for which $|f(x)|$ goes to zero as $x \to 0$ faster than any power of $|x|$ would work.

Given an open interval I and a function $f : I \to \mathbb{R}$, we say that f is **real analytic** at a point $x_0 \in I$ if f is smooth, the formal Taylor series $\sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$ has positive radius of convergence $r = r(x_0) > 0$, and

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

for every x with $|x - x_0| < r$. We say that $f : I \to \mathbb{R}$ is real analytic if it is real analytic at every $x_0 \in I$. Real analytic functions are always smooth, but the example discussed in the previous paragraph shows that not every smooth function is real analytic.

6.2 Formal power series solutions to ODEs

Consider an ODE of the form $f^n = P(t, f, f', \ldots, f^{(n-1)})$ where P is a polynomial in n variables, such as, say,

$$f''' = f^3$$

or

$$f'' = -(f')^2 f^2 + t^5 f.$$

Since we have well-defined notions of differentiation and multiplication for formal power series, we have a well-defined notion of $P(t, f(t), f'(t), \ldots, f^{(n-1)}(t))$ as a formal power series, and hence a well-defined notion of what it means for a formal power series $f(t) \in \mathbb{R}[[t]]$ to satisfy the ODE. We call this a **formal solution** to the ODE. Similarly, we also have a well-defined notion of what it means for a formal power series to satisfy a generalized polynomial ODE of the form $P(t, f, \ldots, f^{(m)}) = 0$, such as $(f')^2 = f$.

Let's see how this works in a very simple example: the ODE $f' = f$. A formal power series $f(t) = \sum_{n=0}^{\infty} a_n t^n \in \mathbb{R}[[t]]$ satisfies this ODE if and only if

$$(n+1)a_{n+1} = a_n$$

for every $n \geq 0$ and, by induction, this holds if and only if

$$a_n = \frac{1}{n!} a_0$$

for every $n \geq 0$. Thus, every formal solution to this ODE is of the form

$$a_0 \sum_{n=0}^{\infty} \frac{t^n}{n!},$$

which is, unsurprisingly, the Taylor series of the exponential.

Of course, once we find a formal solution we would really like to turn it back into a solution in the normal sense (i.e. an honest function). In simple cases we can do this in the direct and obvious way. The **radius of convergence** of the formal power series $\sum_{n=0}^{\infty} a_n x^n$ is defined to be the supremal value of $r \geq 0$ for which the series $\sum_{n=0}^{\infty} |a_n| r^n$ converges. The radius of convergence may be defined in several equivalent ways in addition to this definition:

1. It is equal to $\limsup_{n\to\infty} |a_n|^{1/n}$.
2. It is equal to the supremal value of r for which there exists a constant C_r such that $|a_n| \leq C_r r^{-n}$ for every $n \geq 0$ (i.e., for which $\limsup_{n\to\infty} |a_n| r^n$ is finite).

Here, given a sequence of real numbers c_n, we define $\limsup_{n\to\infty} c_n$ and $\liminf_{n\to\infty} c_n$ by

$$\limsup_{n\to\infty} c_n = \lim_{n\to\infty} \sup_{m\geq n} c_m \qquad \text{and} \qquad \liminf_{n\to\infty} c_n = \lim_{n\to\infty} \inf_{m\geq n} c_m.$$

The limsup and liminf of a sequence always exist as elements of $[-\infty, \infty]$ (since they are limits of monotone sequences) and coincide if and only if $\lim_{n\to\infty} c_n$ is well-defined.

Proposition 6.2 (Formal operations and function operations coincide within radii of convergence)**.** *Let $\sum_{n=0}^{\infty} a_n t^n$ and $\sum_{n=0}^{\infty} b_n t^n$ be formal power series with radii of convergence $r_a, r_b > 0$.*

1. *For each $m \geq 1$, the formal power series $\left(\sum_{n=0}^{\infty} a_n t^n\right)^{(m)} = \sum_{n=0}^{\infty}(n+m)(n+m-1)\cdots(n+1)a_{n+m}t^n$ has radius of convergence r_a. Moreover, the function $f:(-r_a,r_a)\to\mathbb{R}$ defined by $f(t)=\sum_{n=0}^{\infty}a_n t^n$ is smooth with mth derivative*

$$f^{(m)}(t)=\sum_{n=0}^{\infty}(n+m)(n+m-1)\cdots(n+1)a_{n+m}t^n$$

for each $m\geq 1$.

2. *The formal power series $\left(\sum_{n=0}^{\infty}a_n t^n\right)\left(\sum_{n=0}^{\infty}b_n t^n\right)=\sum_{n=0}^{\infty}\left(\sum_{k=0}^{n}a_k b_{n-k}\right)t^n$ has radius of convergence at least $\min\{r_a,r_b\}$. Moreover, if we define functions $f,g,h:(-\min\{r_a,r_b\},\min\{r_a,r_b\})\to\mathbb{R}$ by $f(t)=\sum_{n=0}^{\infty}a_n t^n$, $g(t)=\sum_{n=0}^{\infty}b_n t^n$, and $h(t)=\sum_{n=0}^{\infty}\left(\sum_{k=0}^{n}a_k b_{n-k}\right)t^n$ then $h(t)=f(t)g(t)$ for every $|t|<\min\{r_a,r_b\}$.*

Proof. 1. The fact that the formal mth derivative $\sum_{n=0}^{\infty}(n+m)(n+m-1)\cdots(n+1)a_{n+m}t^n$ has the same radius of convergence as $\sum_{n=0}^{\infty}a_n t^n$ follows since

$$\limsup_{n\to\infty}((n+m)(n+m-1)\cdots(n+1)a_{n+m})^{1/n}=\limsup_{n\to\infty}a_{n+m}^{1/n}=\limsup_{n\to\infty}a_n^{1/n}.$$

The fact that the formal derivative and the actual derivative coincide within the radius of convergence follows by our theorem about differentiating infinite series term by term: If $r<r_a$ and we define functions $f_n:[-r,r]\to\mathbb{R}$ by $f_n(t)=a_n t^n$ then $f_n^{(m)}=n(n-1)\cdots(n-m)a_n t^{n-m}$, so that $\|f_n^{(m)}\|_\infty\leq n(n-1)\cdots(n-m)|a_n|r^{n-m}$ and $\sum_{n=0}^{\infty}\|f_n^{(m)}\|_\infty$ converges by the ratio test.

2. Let r be strictly smaller than the radius of convergence of both $\sum_{n=0}^{\infty}a_n t^n$ and $\sum_{n=0}^{\infty}b_n t^n$, so that there exists a constant C_r such that $|a_n|,|b_n|\leq C_r r^{-n}$ for every $n\geq 0$. This allows us to bound

$$\left|\sum_{k=0}^{n}a_k b_{n-k}\right|\leq nC_r^2 r^{-n}$$

for every $n\geq 0$, from which it follows that the power series $\sum_{n=0}^{\infty}(\sum_{k=0}^{n}a_k b_{n-k})t^n$ has radius of convergence at least r. Since $r<\min\{r_a,r_b\}$ was arbitrary, it follows that the radius of convergence is at least $\min\{r_a,r_b\}$ as claimed.

We have by definition of infinite series that $\sum_{n=0}^{\infty}a_n t^n=\lim_{N\to\infty}\sum_{n=0}^{N}a_n t^n$ and $\sum_{n=0}^{\infty}b_n t^n=\lim_{N\to\infty}\sum_{n=0}^{N}b_n t^n$. Thus, whenever both these limits exist we have that

$$\left(\sum_{n=0}^{\infty}a_n t^n\right)\left(\sum_{n=0}^{\infty}b_n t^n\right)=\lim_{N\to\infty}\left(\sum_{n=0}^{N}a_n t^n\right)\left(\sum_{n=0}^{N}b_n t^n\right).$$

Now, for each N we can write

$$\left(\sum_{n=0}^{N} a_n t^n\right)\left(\sum_{n=0}^{N} b_n t^n\right) = \sum_{n=0}^{2N}\left(\sum_{k=0}^{2N} a_k b_{n-k}\mathbb{1}(k, n-k \le N)\right) t^n$$
$$= \sum_{n=0}^{N}\left(\sum_{k=0}^{n} a_k b_{n-k}\right) t^n + \sum_{n=N+1}^{2N}\left(\sum_{k=0}^{n} a_k b_{n-k}\mathbb{1}(k, n-k \le N)\right) t^n.$$

Letting r be strictly smaller than the radius of convergence of both $\sum_{n=0}^{\infty} a_n t^n$ and $\sum_{n=0}^{\infty} b_n t^n$ and using the bound $|a_n|, |b_n| \le C_r r^{-n}$ as above allows us to bound the error term appearing here

$$\left|\sum_{n=N+1}^{2N}\left(\sum_{k=0}^{n} a_k b_{n-k}\mathbb{1}(k, n-k \le N)\right) t^n\right| \le \sum_{n=N+1} n C_r^2 (t/r)^n,$$

so that

$$\left|\sum_{n=N+1}^{2N}\left(\sum_{k=0}^{n} a_k b_{n-k}\mathbb{1}(k, n-k \le N)\right) t^n\right| \to 0 \qquad \text{as } N \to \infty \text{ when } t < r.$$

Thus, it follows that if $|t| < r$ then

$$\left(\sum_{n=0}^{\infty} a_n t^n\right)\left(\sum_{n=0}^{\infty} b_n t^n\right) = \lim_{N\to\infty}\left(\sum_{n=0}^{N} a_n t^n\right)\left(\sum_{n=0}^{N} b_n t^n\right)$$
$$= \lim_{N\to\infty}\sum_{n=0}^{N}\left(\sum_{k=0}^{n} a_k b_{n-k}\right) t^n + \sum_{n=N+1}^{2N}\left(\sum_{k=0}^{n} a_k b_{n-k}\mathbb{1}(k, n-k \le N)\right) t^n$$
$$= \sum_{n=0}^{N}\left(\sum_{k=0}^{n} a_k b_{n-k}\right) t^n = \sum_{n=0}^{\infty}\sum_{k=0}^{n} a_k b_{n-k} t^n,$$

and the claim follows since $r < \min\{r_a, r_b\}$ was arbitrary. □

This proposition has the following consequence: If $\sum_{n=0} a_n t^n$ is a formal power series solution to a generalized polynomial ODE $P(t, f, \ldots, f^{(n)}) = 0$ that has positive radius of convergence, then the honest function defined through this power series is a solution to the ODE within that radius of convergence. Moreover, in this situation the derivatives of f at zero are all given by $f^{(n)}(0) = n! a_n$.

In the example $f' = f$ we studied above, the formal solution $a_0 \sum_{n=0}^{\infty} \frac{t^n}{n!}$ has infinite radius of convergence and we deduce that the function $\mathbb{R} \to \mathbb{R}$ defined by this power series (namely, the exponential function) is an honest solution to the ODE.

We can also go in the other direction:

Corollary 6.3. *If I is an open interval containing 0 and (I, f) is a solution to the generalized polynomial ODE $P(t, f, \ldots, f^{(m)}) = 0$ that is real analytic at zero, then the Taylor series $\sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} t^n$ is a formal solution to the same ODE.*

Example 6.4. Let's now consider the ODE $f' = f^2$. We already studied this equation using other methods and saw that every maximal solution is either of the form $(\mathbb{R}, 0)$, $((c, \infty), 1/(c-t))$, or $((-\infty, c), 1/(c-t))$. For a formal power series $\sum_{n=0}^{\infty} a_n t^n$ to be a formal solution to the ODE, we must have that

$$\sum_{n=0}^{\infty} (n+1) a_{n+1} t^n = \left(\sum_{n=0}^{\infty} a_n t^n \right)' = \left(\sum_{n=0}^{\infty} a_n t^n \right)^2 = \sum_{n=0}^{\infty} \left(\sum_{k=0}^{n} a_k a_{n-k} \right) t^n.$$

By definition, this means that the equality

$$(n+1) a_{n+1} = \sum_{k=0}^{n} a_k a_{n-k}$$

holds for every $n \geq 0$, and hence that

$$a_{n+1} = \frac{1}{n+1} \sum_{k=0}^{n} a_k a_{n-k}$$

for every $n \geq 0$. Note that this uniquely specifies the whole sequence $(a_n)_{n \geq 0}$ in terms of a_0. To find a formula for general n, we can input a few small values, guess a formula, then prove this formula by induction. To this end, we see that

$$a_1 = a_0^2, \qquad a_2 = \frac{1}{2}(a_0 a_1 + a_1 a_0) = a_0^3, \qquad a_3 = \frac{1}{3}(a_0 a_2 + a_1 a_1 + a_2 a_0) = a_0^4, \qquad \ldots$$

and it seems reasonable to guess that $a_n = a_0^{n+1}$ for every $n \geq 1$. It is easy to check by induction that this does indeed work: If $a_k = a_0^{k+1}$ for every $0 \leq k \leq n$ then

$$a_{n+1} = \frac{1}{n+1} \sum_{k=0}^{n} a_k a_{n-k} = \frac{1}{n+1} \sum_{k=0}^{n} a_0^{n+2} = a_0^{n+2}$$

as required. This means that every formal solution to our ODE is of the form

$$\sum_{n=0}^{\infty} a_0^{n+1} t^n$$

for some $a_0 \in \mathbb{R}$. For each $a_0 \neq 0$ this series has radius of convergence $|a_0|^{-1}$ and, as expected, the function defined by summing the series within this radius of convergence coincides with the function $a_0/(1 - a_0 t) = 1/(a_0^{-1} - t)$ within this radius of convergence.

As such, we see in this example that: 1) the formal solution can be summed within its radius of convergence to give us a solution to the ODE, but this solution is not necessarily maximal. 2) We naturally do not see the maximal solutions that are not defined at zero as formal power series solutions.

Remark 6.5. In general when we find the formal solution to an ODE there is no reason to expect that we can find a nice formula for the coefficients or that we can recognize the resulting function as 'something we know'. (Of course this is not a well-defined mathematical notion!)

Example 6.6. Consider the linear ODE

$$f'' = tf' + f,$$

which we previously studied using the Laplace transform. For a formal power series $\sum_{n=0}^{\infty} a_n t^n$ to be a formal solution to the ODE, we must have that

$$\begin{aligned}\sum_{n=0}^{\infty}(n+2)(n+1)a_{n+2}t^n &= t\sum_{n=0}^{\infty}(n+1)a_{n+1}t^n + \sum_{n=0}^{\infty}a_n t^n \\ &= \sum_{n=0}^{\infty} n\mathbb{1}(n\geq 1)a_n t^n + \sum_{n=0}^{\infty}a_n t^n = \sum_{n=0}^{\infty}(n+1)a_n t^n\end{aligned}$$

By definition, this means that the equality

$$(n+2)(n+1)a_{n+2} = (n+1)a_n$$

holds for every $n \geq 0$. By induction, this means that

$$a_n = \begin{cases} \left(\prod_{k=0}^{n/2-1}(n-2k)\right)^{-1} a_0 & n \text{ even} \\ \left(\prod_{k=0}^{\lfloor n/2\rfloor}(n-2k)\right)^{-1} a_1 & n \text{ odd} \end{cases}.$$

The products appearing here, which is equal to the product of all positive numbers smaller than n with the same parity as n, is known (for some reason) as the **double factorial** and denoted by $n!!$. (Note that his is *not* the same thing as $(n!)!$, which is **much** larger than $n!!$.) Using this notation allows us to write our formal solution neatly as

$$\sum_{n=0}^{\infty}\frac{a_0\mathbb{1}(n \text{ even}) + a_1\mathbb{1}(n \text{ odd})}{n!!}t^n.$$

This formal power series has infinite radius of convergence by the ratio test. Thus, any function defined by one of these formal power series is a maximal solution to the ODE with $f(0) = a_0$ and $f'(0) = a_1$; it follows from global Picard-Lindelöf that *every* maximal solution is of this form. On the other hand, this series doesn't look like the Taylor series of anything we're familiar with, so we don't obviously obtain a closed-form solution (i.e., a solution in

terms of compositions of known functions). This isn't necessarily a problem depending on what we want to do with our solution, but in fact in this case we can write the solution in terms of standard functions with some work.

Indeed, note that if $a_1 = 0$ then, since $(2m)!! = 2^m m!$ for every $m \geq 0$, our solution is given by

$$\sum_{n=0}^{\infty} \frac{a_0}{2^n n!} t^{2n} = a_0 e^{\frac{1}{2}t^2}.$$

Double factorials of odd numbers are not so nice, so it's unclear what we would do when a_1 is not zero. In the following exercise you will give a solution by quadrature using the method of *Wronskians*, which, in general, lets us compute an nth linearly independent solution of an nth order linear ODE in terms of (integrals of) any collection of $n-1$ linearly independent solutions.

Exercise 53 (Wronskians)**.** Suppose that $f, g : \mathbb{R} \to \mathbb{R}$ are two solutions to the second-order homogeneous linear ODE $f'' + a(t)f' + b(t)f = 0$, with $a, b : \mathbb{R} \to \mathbb{R}$ continuous. The **Wronskian** of f and g is the function $W(f, g) : \mathbb{R} \to \mathbb{R}$ defined by

$$W(f, g) = \det \begin{pmatrix} f & g \\ f' & g' \end{pmatrix} = fg' - f'g.$$

1. Prove that if $W(f, g)(t_0) \neq 0$ for some t_0 then $\{f, g\}$ is a basis for the space of solutions to the ODE.

2. Prove that $W(f, g)$ satisfies the ODE $W(f, g)' = -a(t)W(f, g)$.

3. Find a first-order linear ODE satisfied by g defined in terms of f and $W(f, g)(0)$, and use this to give a formula for g in terms of f, a, b, $g(0)$ and $g'(0)$.

4. Find a basis of solutions to the ODE $f'' = tf' + f$. [You may use that $e^{t^2/2}$ is a solution.]

Example 6.7. Consider the ODE $tf' - 2f - 2t^2 = 0$. This ODE is **not** in our usual form, since the top derivative f' has not been solved for, but it is of the form $P(t, f, f') = 0$ for a polynomial P so that there is a well-defined notion of what is means for a formal power series to be a formal solution. Indeed, $\sum_{n=0}^{\infty} a_n t^n$ is a formal solution if and only if

$$\sum_{n=0}^{\infty} \mathbb{1}(n \geq 1) n a_n t^n - 2t^2 - \sum_{n=0}^{\infty} 2a_n t^n.$$

In order for this to hold we must have that $2a_2 - 2 - 2a_2 = 0$, which is impossible, so that there do not exist any formal power series solutions to this ODE. On the other hand, we can check that

$$f(t) = \begin{cases} t^2 \log t^2 & t \neq 0 \\ 0 & t = 0 \end{cases}$$

is a solution to the ODE. This function is not analytic at zero (indeed, it is not three-times differentiable at zero).

Theorem 6.8 (Existence and uniqueness of formal solutions)**.** *Every polynomial ODE of the form $f^{(m)} = P(t, f, \dots, f^{(m-1)})$ has exactly one formal solution $f(t) = \sum_{n=0}^{\infty} a_n t^n$ for each given values of $a_0, \dots, a_{n-1}$.*

Proof. If $f(t) = \sum_{n=0}^{\infty} a_n t^n$ then the formal power series $P(t, f, \dots, f^{(m-1)})$ has the property that the kth coefficient $P(t, f, \dots, f^{(m-1)})_k$ is determined by the first $k + m - 1$ coefficients $a_0, \dots, a_{k+m-1}$ of f for each $k \geq 0$. This can be verified formally by induction on the degree of P, using that the first k coefficients of $f^{(m)}$ are determined by the first $k + m$ coefficients of f, that the first k coefficients of the product $(\sum_{n=0}^{\infty} a_n t^n)(\sum_{n=0}^{\infty} b_n t^n)$ are determined by the first k coefficients of $\sum_{n=0}^{\infty} a_n t^n$ and $\sum_{n=0}^{\infty} b_n t^n$, and that the first k coefficients of $t^r \sum_{n=0}^{\infty} a_n t^n$ are determined by the first $k - r$ coefficients of $\sum_{n=0}^{\infty} a_n t^n$. Thus, the ODE $f^{(m)} = P(t, f, \dots, f^{(m-1)})$ determines the $n + m$th coefficient of f as a function of the first $n+m-1$ coefficients of f for each $n \geq 0$, and it follows that there is a unique formal solution for each given value of $(a_0, \dots, a_{m-1})$. □

Theorem 6.9 (Positive radius of convergence)**.** *Let $f(t) = \sum_{n=0}^{\infty} a_n t^n$ be a formal solution of a polynomial ODE of the form $f^{(m)} = P(t, f, \dots, f^{(m-1)})$. Then $\sum_{n=0}^{\infty} a_n t^n$ has positive radius of convergence.*

This theorem together with the equivalence of formal and function operations within the radius of convergence yields the following corollary, which is a special case of *Cauchy's Theorem.*

Corollary 6.10. *If $I \subseteq \mathbb{R}$ is open and (I, f) is a solution to the polynomial ODE $f^{(m)} = P(t, f, \dots, f^{(m-1)})$ then f is real analytic on I.*

Cauchy's Theorem states more generally that solutions to ODEs of the form $f^{(m)} = F(t, f, \dots, f^{(m-1)})$ where F is analytic are analytic (see also the Cauchy–Kovalevskaya theorem for PDEs). Cauchy's theorem is usually proven using sophisticated tools from complex analysis etc. We will prove Theorem 6.9 in a more direct and elementary way by recursively bounding the coefficients of the solution. (Warning: "Elementary" proofs can be significantly more complicated than their more conceptually sophisticated counterparts!)

(This proof was not lectured in 2024.)

Proof of Theorem 6.9. Let $a_0, a_1, \dots, a_{m-1}$ be given and let the coefficients $(a_n)_{n \geq m}$ be determined by letting $f(t) = \sum_{n=0}^{\infty} a_n t^n$ be a formal solution to the ODE $f^{(m)} = P(t, f, \dots, f^{(m-1)})$. Express $P(t, f, \dots, f^{(m-1)})$ as

$$P(t, f, \dots, f^{(m-1)}) = \sum_{i=1}^{\ell} \lambda_i t^{r_i} f^{k_{0,i}} \cdots (f^{(m-1)})^{k_{m-1,i}}$$

for some real numbers $\lambda_1, \dots, \lambda_\ell$ and non-negative integers r_i and $k_{j,i}$ with $1 \leq i \leq \ell$ and $0 \leq j \leq m-1$, let $\tilde{P}$ be the polynomial defined by

$$\tilde{P}(t, f, \dots, f^{(m-1)}) = f^{(m-1)} + \sum_{i=1}^{\ell} |\lambda_i| t^{r_i} f^{k_{0,i}} \cdots (f^{(m-1)})^{k_{m-1,i}},$$

so that all the coefficients of $\tilde{P}$ are non-negative. Let the sequence $(\tilde{a}_n)_{n\geq 0}$ be given by taking $\tilde{a}_n = 1 + \max_{0\leq i\leq m-1} |a_i|$ for every $0 \leq n \leq m-1$ and letting $(\tilde{a}_n)_{n\geq m}$ be determined by letting $\tilde{f}(t) = \sum_{n=0}^{\infty} \tilde{a}_n t^n$ be a formal solution to the ODE $\tilde{f}^{(m)} = \tilde{P}(t, \tilde{f}, \dots, \tilde{f}^{(m-1)})$, so that the sequence $(\tilde{a}_n)_{n\geq 0}$ is non-negative. Since $|a * b|_n \leq (|a| * |b|)_n$ for any two sequences $a = (a_n)_{n\geq 0}$ and $b = (b_n)_{n\geq 0}$, we have by induction on the degree of P that

$$|a_n| \leq \tilde{a}_n$$

for every $n \geq 0$. Moreover, the inclusion of the f term in the definition of $\tilde{P}$, together with the fact that all the coefficients of $\tilde{P}$ are non-negative, ensures that the sequence $(\tilde{a}_n)_{n\geq 0}$ satisfies the inequality

$$(n+m)(n+m-1)\cdots(n+1)\tilde{a}_{n+m} \geq (n+m-1)\cdots(n+1)\tilde{a}_{n+m-1}$$

for every $n \geq 0$, and hence that

$$(n+1)\tilde{a}_{n+1} \geq \tilde{a}_n$$

for every $n \geq 0$, where the fact that this holds for $0 \leq n \leq m-1$ follows by definition of $\tilde{a}_0, \dots, \tilde{a}_{m-1}$. Applying this inequality recursively we deduce that

$$(n+k)(n+k-1)\cdots(n+1)\tilde{a}_{n+k} \geq (n+k-1)\cdots(n+1)\tilde{a}_{n+k-1} \geq \cdots \geq \tilde{a}_n$$

for every $n, k \geq 0$. In other words, if $\tilde{f}(t) = \sum_{n=0} \tilde{a}_n t^n$ denotes the formal power series determined by $(\tilde{a}_n)_{n\geq 0}$ then the nth coefficient of the formal derivative $\tilde{f}(t)^{(k)}$ is an increasing function of k for each $n \geq 0$, and hence that

$$\begin{aligned}\tilde{P}(t, \tilde{f}, \dots, \tilde{f}^{(m-1)})_n &= \left(\tilde{f}^{(m-1)} + \sum_{i=1}^{\ell} |\lambda_i| t^{r_i} \tilde{f}^{k_{0,i}} \cdots (\tilde{f}^{(m-1)})^{k_{m-1,i}} \right)_n \\ &\leq \left(\tilde{f}^{(m-1)} + \sum_{i=1}^{\ell} |\lambda_i| t^{r_i} (\tilde{f}^{(m-1)})^{k_i} \right)_n\end{aligned}$$

where $k_i = \sum_{j=0}^{m-1} k_{j,i}$.

Our aim is to bound the coefficients $\tilde{a}_n$ in terms of the coefficients of the formal solution to a simpler ODE that we can solve exactly. To this end, let $\lambda = 1 + \sum_{i=1}^{\ell} |\lambda_i|$, let $k = 2 + \max_{1\leq i\leq \ell} k_i$, and consider the ODE $g' = \lambda g^k$. We can solve the ODE $g' = \lambda g^k$ non-

formally by thinking of it as a separable equation to find the family of solutions

$$g = \frac{\lambda^{1/(k-1)}(k-1)^{1/(k-1)}}{(t_0-t)^{1/(k-1)}} \qquad t_0 > 0, \quad t < t_0,$$

and the generalized binomial theorem allows us to write this solution as a convergent power series

$$\begin{aligned}\frac{\lambda^{1/(k-1)}(k-1)^{1/(k-1)}}{(t_0-t)^{1/(k-1)}} &= \frac{\lambda^{1/(k-1)}(k-1)^{1/(k-1)}}{t_0^{1/(k-1)}}(1-(t/t_0))^{-1/(k-1)} \\ &= \left(\frac{\lambda(k-1)}{t_0}\right)^{1/(k-1)} \sum_{n=0}^{\infty} \left[\frac{(-1)^n}{t_0^n n!}\prod_{i=0}^{n-1}(-i-1/(k-1))\right] t^n.\end{aligned}$$

and it follows that

$$g(t) = \sum_{n=0}^{\infty} \left(\frac{\lambda(k-1)}{t_0}\right)^{1/(k-1)} \left[\frac{(-1)^n}{t_0^n n!}\prod_{i=0}^{n-1}(-i-1/(k-1))\right] t^n.$$

is also a *formal* solution to the ODE $g' = \lambda g^k$ for each $t_0 > 0$. The coefficients of this power series are all positive since the two negatives always cancel for n odd, and since they become large when t_0 is small we can take t_0 sufficiently small that these coefficients satisfy $g_n \geq \tilde{f}_n^{(m-1)}$ for every $0 \leq n \leq m$. Moreover, we also have that

$$\frac{g_{n+1}}{g_n} = \frac{(n+1/(k-1))}{(n+1)t_0}$$

for every $n \geq 0$, so that if t_0 is sufficiently small then $g_{n+1} \geq g_n$ for every $n \geq 0$. This implies that the coefficients of g^ℓ are also increasing for every $\ell \geq 1$:

Exercise 54. Prove that if $\sum_{n=0}^{\infty} a_n t^n$ and $\sum_{n=0}^{\infty} b_n t^n$ are formal power series such that the sequences $(a_n)_{n\geq 0}$ and $(b_n)_{n\geq 0}$ are both non-negative and increasing then the coefficients of the product $(\sum_{n=0}^{\infty} a_n t^n)(\sum_{n=0}^{\infty} b_n t^n)$ are non-negative and increasing. Deduce that $(\sum_{n=0}^{\infty} a_n t^n)^\ell$ has non-negative, increasing coefficients for every $\ell \geq 1$.

Exercise 55. Prove that if $f(t) = \sum_{n=0}^{\infty} a_n t^n$ is a formal power series with $a_0 \geq 1$ and the sequence (a_n) is non-negative then $f_n^\ell \geq f_n$ for every $n \geq 0$ and $\ell \geq 1$.

We claim that if t_0 is small enough that both inequalities hold then $g_n \geq \tilde{f}_n^{(m-1)}$ for *every* $n \geq 0$. We prove this by induction on n, the base case $n = 0$ being trivial by choice of t_0. Suppose that the claim has been proven for all $0 \leq i \leq n$ for some $n \geq 0$. Then we have that

$$\tilde{f}_{n+1}^{(m-1)} = \frac{1}{n+1}\tilde{f}_n^{(m)} = \frac{1}{n+1}\tilde{P}(t, \tilde{f}, \ldots, \tilde{f}^{(m-1)})_n \leq \frac{1}{n+1}\left(\tilde{f}^{(m-1)} + \sum_{i=1}^{\ell} |\lambda_i| t^{r_i} (\tilde{f}^{(m-1)})^{k_i}\right)_n.$$

Since the right hand side is an increasing function of the coefficients $f_0^{(m-1)}, \dots, f_n^{(m-1)}$, it follows from the induction hypothesis that

$$\tilde{f}_{n+1}^{(m-1)} \leq \frac{1}{n+1}\left(g + \sum_{i=1}^{\ell} |\lambda_i| t^{r_i} g^{k_i}\right)_n \leq \frac{\lambda g_n^k}{n+1} = \frac{g_n'}{n+1} = g_{n+1}$$

where the second inequality holds by our two exercises and the penultimate equality follows since g is a formal solution to the ODE $g' = \lambda g^k$.

Putting this all together, it follows that

$$|a_n| \leq \tilde{a}_n = \frac{1}{n(n-1)\cdots(n-m+2)} \tilde{f}_{n-m+1}^{(m-1)} \leq g_{n-m+1}$$

for every $n \geq m$, and since g has positive radius of convergence it follows that $\sum_{n=0}^{\infty} a_n t^n$ does also. □

Example 6.11. Consider the ODE $\frac{t^2}{2} f' + \frac{t^2}{2} - f = 0$. This ODE is **not** in our usual form, since the top derivative f' has not been solved for, but it is of the form $P(t, f, f') = 0$ for a polynomial P so that there is a well-defined notion of what is means for a formal power series to be a formal solution. Indeed, $\sum_{n=0}^{\infty} a_n t^n$ is a formal solution if and only if

$$\sum_{n=0}^{\infty} \mathbb{1}(n \geq 2)(n-1)\frac{a_{n-1}}{2} t^n + \frac{t^2}{2} - \sum_{n=0}^{\infty} a_n t^n.$$

For this to hold we must have that $a_0 = a_1 = 0$, that $a_2 = \frac{1}{2}$, and that

$$\frac{(n-1)a_{n-1}}{2} = a_n$$

for every $n \geq 3$, so that, inductively,

$$a_n = \frac{(n-1)}{2} a_{n-1} = \frac{(n-1)}{2}\frac{(n-2)}{2} a_{n-2} = \cdots = \frac{(n-1)!}{2^{n-2}} a_2 = \frac{(n-1)!}{2^{n-1}}$$

for every $n \geq 2$. Thus, the *only* formal solution to the ODE is given by

$$\sum_{n=0}^{\infty} \frac{(n-1)!}{2^{n-1}} \mathbb{1}(n \geq 2) t^n,$$

which has zero radius of convergence. Later we will see that this formal solution is still meaningful even though it cannot be summed to give a function solving the ODE.

Exercise 56. Find all formal power series solutions to Stokes equation $f'' = tf$.

Exercise 57. Find all formal power series solutions to $f'' = f'f$.

6.3 Composition of formal power series and analytic ODEs

So far we have been discussing formal power series solutions only in the context of *polynomial* ODEs. While this is already a very large class (including most examples one encounters in practice), the theory can also be applied to *analytic* ODEs using the notion of the formal composition of formal power series.

Suppose that $f(t) = \sum_{n=0}^{\infty} a_n t^n$ and $g(t) = \sum_{n=0}^{\infty} b_n t^n$. We would like to be able to define the formal composition by

$$f(g(t)) = \sum_{n=0}^{\infty} a_n g(t)^n = a_0 + a_1(b_0 + b_1 t + \cdots) + a_2(b_0 + b_1 t + \cdots)^2 + \cdots .$$

If b_0 is not zero, then we get infinitely many constant terms when expanding the right hand side, so that the series cannot be defined without getting into issues of convergence – which we do not want to do when using formal power series! On the other hand, **if the constant term b_0 is equal to zero**, we only get finitely many terms contributing to the coefficient of each power of t and can therefore **define** the formal composition $f(g(t))$ by

$$f(g(t)) = a_0 + \sum_{n=0}^{\infty} \left(\sum_{k=1}^{n} a_k \sum_{\substack{j_1,\ldots,j_k \geq 1 \\ \sum j_i = n}} b_{j_1} \cdots b_{j_k} \right) t^n.$$

Similarly, if $f : I \to \mathbb{R}$ is a real analytic function defined on some open set I and $g(t) = \sum_{n=0}^{\infty} b_n t^n$ is a formal power series with $b_0 \in I$, we can write f as a power series around b_0

$$f(t) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a_0)}{n!}(t - b_0)^n$$

and define the formal composition $f(g(t))$ to be the formal power series

$$\begin{aligned} f(g(t)) &= f(b_0) + f'(b_0)(b_1 t + b_2 t^2 + \cdots) + \frac{f''(b_0)}{2!}(b_1 t + b_2 t^2 + \cdots)^2 + \cdots \\ &= f(b_0) + \sum_{n=0}^{\infty} \left(\sum_{k=1}^{n} \frac{f^{(k)}(b_0)}{k!} \sum_{\substack{j_1,\ldots,j_k \geq 1 \\ \sum j_i = n}} b_{j_1} \cdots b_{j_k} \right) t^n. \end{aligned}$$

Thus, there is a well-defined notion of a formal power series $f(t) = \sum_{n=0}^{\infty} a_n t^n$ being a formal solution to an ODE of the form, say, $f^{(m)} = F(f)$ where F is a real analytic function whose domain contains a_0. (Of course one can also define formal solutions to ODEs of the form $F(t, f, \ldots, f^{(m)}) = 0$ for F analytic, but I don't want to get into a discussion of analytic functions in multiple variables.)

Example 6.12. Let $f(t) = \sum_{n=0}^{\infty} a_n t^n$ be a formal solution to the ODE $f' = \cos(f)$ with $a_0 = 0$. This ODE means that

$$\sum_{n=0}^{\infty} a_n t^n = 1 + \sum_{n=0}^{\infty} \left(\sum_{k=1}^{n} \frac{(-1)^{k/2}}{k!} \mathbb{1}(k \text{ even}) \sum_{\substack{j_1,\ldots,j_k \geq 1 \\ \sum j_i = n}} a_{j_1} \cdots a_{j_k} \right) t^n.$$

We can solve for the first few coefficients by hand:

$$\begin{aligned}
a_1 &= 1 \\
2a_2 &= 0 \\
3a_3 &= -\frac{1}{2}a_1^2 && \Rightarrow a_3 = -\frac{1}{6} \\
4a_4 &= 0 \\
5a_5 &= -\frac{1}{2}a_1 a_3 - \frac{1}{2}a_3 a_1 + \frac{1}{4!}a_1^4 && \Rightarrow a_5 = \frac{1}{24}.
\end{aligned}$$

One can take the same calculations further with a computer to obtain that

$$f(t) = t - \frac{1}{6}t^3 + \frac{1}{24}t^5 - \frac{61}{5040}t^7 + \frac{277}{72576}t^9 - \cdots$$

It is easy to show by induction that $a_n = 0$ for every even n, but no clear pattern emerges for the odd coefficients. This is just what is to be expected when solving formal ODEs in general!

On the other hand, the ODE $f' = \cos(f)$ can also be solved as a separable ODE to obtain the (function) solution

$$f(t) = 2\tan^{-1}\left(\tanh\left(\frac{1}{2}x\right)\right),$$

which is sometimes known[13] as the **Gudermannian function**. The derivatives of this function at zero are complicated, which is why we didn't see any obvious pattern in our formal power series solution.

[13] I had never heard of this function before I prepared this example.

7 Recursions, difference equations, and generating functions

Series solutions allow us to reduce the problem of finding the solutions to a polynomial ODE $f^{(m)} = P(t, f, \dots, f^{(m-1)})$ to the problem of recursively computing the coefficients $a_0, a_1, \dots$ of $\sum_{n=0}^{\infty} a_n t^n$ via the relation

$$a_{n+m} = \frac{1}{(n+m)(n+m-1)\cdots(n+1)} P(t, f, \dots, f^{(m-1)})_n$$

where the right hand side is a function of the coefficients $a_0, \dots, a_{n+m-1}$. In particular, this method allows us to turn a *continuous problem* (solving an ODE) into a *discrete problem* (computing a recursively defined sequence). In this section we discuss how this method can also be used in *reverse*, to study discrete problems about recursively defined sequences using ODE methods.

7.1 Ordinary generating functions

Given a sequence of real numbers $(a_n)_{n\geq 0}$, the **formal ordinary generating function** of $(a_n)_{n\geq 0}$ is the formal power series $\sum_{n=0}^{\infty} a_n t^n$. The **ordinary generating function** of $(a_n)_{n\geq 0}$ is the function defined by summing this formal power series within its radius of convergence (taking the generating function to have empty domain when the radius of convergence is zero). The word "ordinary" is used to distinguish the ordinary generating function from the *exponential* generating function, which we will discuss later.

The operation of sending a sequence to its generating function is closely analogous to the operation of sending a function to its Laplace transform. Indeed, if we define a piecewise-continuous function g by $g(t) = a_n$ for every $n \geq 0$ and every $n \leq t < n+1$ then

$$\mathcal{L}\{g\}(s) = \sum_{n=0}^{\infty} \int_n^{n+1} a_n e^{-st} \, \mathrm{d}t = \frac{1-e^{-s}}{s} \sum_{n=0}^{\infty} a_n e^{-sn}$$

whenever all relevant series and integrals converge absolutely, so that the ordinary generating function of $(a_n)_{n\geq 0}$ and the Laplace transform of g are 'the same' up to a change in parameterization and multiplication by a simple $\frac{1-e^{-s}}{s}$ prefactor.

As with the Laplace transform, there are many identities relating operations on sequences to operations on ordinary generating functions. Since we are doing everything formally, all these rules follow directly from the definitions and we do not need to be careful about convergence issues as we did when discussing the Laplace transform.

1. (Shifting index $\longrightarrow$ multiplication by $1/t$.) If $f(t)$ is the formal ordinary generating function of $(a_n)_{n\geq 0}$ then $(f(t) - a_0)/t$ is the formal ordinary generating function of

$(a_{n+1})_{n\geq 0}$. Similarly, if $k \geq 1$ then

$$\frac{f(t) - a_0 - a_1 t - \cdots - a_{k-1}t^{k-1}}{t^k}$$

is the formal ordinary generating function of $(a_{n+k})_{n\geq 0}$.

2. (Multiplication by $n \longrightarrow$ multiplication by $t\frac{d}{dt}$.) If $f(t)$ is the formal ordinary generating function of $(a_n)_{n\geq 0}$ then $tf'(t)$ is the formal ordinary generating function of $(na_n)_{n\geq 0}$. More generally, if P is a real polynomial in one variable then the ordinary generating function of $(P(n)a_n)_{n\geq 0}$ is given by $P(t\frac{d}{dt})f(t)$.

3. (Convolutions $\longrightarrow$ products) If $f(t)$ and $g(t)$ are the formal ordinary generating functions of $a = (a_n)_{n\geq 0}$ and $b = (b_n)_{n\geq 0}$ respectively then $f(t)g(t)$ is the formal ordinary generating function of $a * b$. In particular, f^k is the formal ordinary generating function of the sequence

$$\left(\sum_{\substack{n_1,\ldots n_k \geq 0 \\ \sum n_i = n}} \prod_{i=1}^{k} a_{n_i} \right)_{n\geq 0}.$$

Let us now go through some simple examples where we can use ordinary generating functions to solve recurrences.

Example 7.1. Let $N(n,k)$ be the number of ways of writing n as the ordered sum of k non-negative integers. For example, $N(2,2) = 3$ since we can write 2 as $2+0$, $1+1$, and $0+2$. If we let $a = (a_n)_{n\geq 0}$ be the constant sequence $a_n \equiv 1$, then

$$N(n,k) = \sum_{\substack{n_1,\ldots n_k \geq 0 \\ \sum n_i = n}} 1 = \sum_{\substack{n_1,\ldots n_k \geq 0 \\ \sum n_i = n}} \prod_{i=1}^{k} a_{n_i}$$

so that $N(n,k)$ is the nth coefficient of the formal power series $(\sum_{n=0}^{\infty} t^n)^k$. This formal power series can be summed to obtain the function

$$\left(\sum_{n=0}^{\infty} t^n \right)^k = \frac{1}{(1-t)^k}$$

for $|t| < 1$. The generalized binomial theorem lets us expand this series as

$$\frac{1}{(1-t)^k} = \sum_{n=0}^{\infty} \binom{n+k-1}{n} t^n,$$

so that $N(n,k) = \binom{n+k-1}{n}$ for every $n, k \geq 1$.

Exercise 58. Find a formula for the number of ways of writing n as the ordered sum of k *positive* integers.

Example 7.2 (Catalan numbers)**.** For each $n \geq 0$ let C_n be the number of ways of writing n pairs of left and right parentheses so that every left parenthesis is correctly matched to a right parenthesis (i.e., such that when we read the sequence from left to right, we have at each step read at least as many left parenthesis as right parentheses). For example,

$$C_1 = \#\{()\} = 1, \quad C_2 = \#\{()(), (())\} = 2, \quad C_3 = \#\{()()(), ()(()), (())(), (()()), ((()))\} = 5.$$

To make things work nicely, we also consider the empty sequence, with zero pairs of parentheses, to be a sequence of parentheses with all parentheses correctly matched and set $C_0 = 1$. C_n is known as the nth **Catalan number** and arises in a huge variety of combinatorial enumeration problems (see e.g. Figure 6). In any such sequence of parentheses, the first parenthesis must be a left parenthesis, which is matched to some right parenthesis, so that we can write our sequence uniquely in the form $(X)Y$ where X and Y are (possibly length zero) sequences of correctly-matched parentheses. Considering the possible lengths of the two sequences X and Y leads to the recurrence

$$C_{n+1} = \sum_{k=0}^{n} C_k C_{n-k} = (C * C)_n$$

which holds for every $n \geq 0$. Thus, if $f(t) = \sum_{n=0}^{\infty} C_n t^n$ is the formal ordinary generating function of $(C_n)_{n\geq 0}$ then

$$\frac{f-1}{t} = f^2.$$

If $\sum_{n=0}^{\infty} C_n t^n$ has positive radius of convergence then f must satisfy this equality *as a function* within this radius of convergence, and we obtain that

$$f = \frac{1+\sqrt{1-4t}}{2t} \qquad \text{or} \qquad f = \frac{1-\sqrt{1-4t}}{2t}$$

for every positive t within the radius of convergence. The first option is not viable since it converges to ∞ as $t \downarrow 0$. L'Hopital's rule implies that the second expression converges to 1 as $t \downarrow 0$, and expanding $\sqrt{1-4t}$ as a power series using the generalized binomial theorem yields that

$$\begin{aligned}\sqrt{1-4t} &= \sum_{n=0}^{\infty} \binom{1/2}{n} (-4t)^n = 1 + \sum_{n=0}^{\infty} \frac{(\frac{1}{2})(\frac{1}{2}-1)\cdots(\frac{1}{2}-n+1)}{n!}(-4t)^n \\ &= 1 - \sum_{n=1}^{\infty} \frac{2^n(2n-3)!!}{n!} t^n,\end{aligned}$$

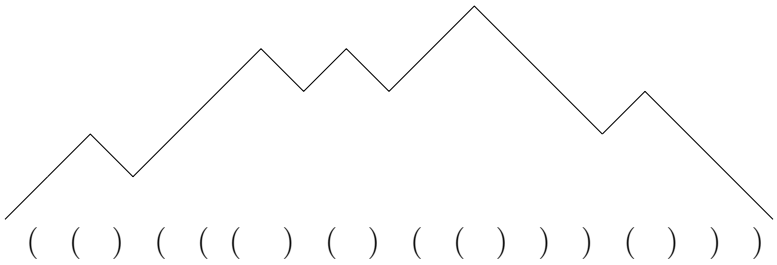


Figure 6: The nth Catalan number C_n is also equal to the number of sequences $(X_0, X_1 \dots X_{2n})$ such that $|X_i - X_{i-1}| = \pm 1$ for each $1 \leq i \leq 2n$, $X_0 = X_{2n} = 0$, and $X_i \geq 0$ for every $0 \leq i \leq 2n$. Such sequences are called **Dyck paths**. To see this, note that there is a bijection between such paths and configurations of parentheses defined by taking $X_i - X_{i-1} = 1$ if the ith parenthesis is a left parenthesis and $X_i - X_{i-1} = -1$ if the ith parenthesis is a right parenthesis. As such, $2^{-n}C_n$ is equal to the probability that a $2n$-step simple random walk on the integers with $X_0 = 0$ satisfies $X_{2n} = 0$ and $X_i \geq 0$ for every $0 \leq i \leq 2n$.

where we use the convention that $(-\ell)!! = 1$ for $\ell \geq 0$, so that

$$\frac{1 - \sqrt{1-4t}}{2t} = \frac{1}{2t} \sum_{n=1}^{\infty} \frac{2^n (2n-3)!!}{n!} t^n = \sum_{n=0}^{\infty} \frac{2^n (2n-1)!!}{(n+1)!} t^n.$$

Since this power series has positive radius of convergence, satisfies the equation $\frac{f-1}{t} = f^2$, and takes the value 1 at 0, the coefficients $a_n = \frac{2^n(2n-1)!!}{(n+1)!}$ must satisfy the recurrence $a_0 = 1$, $a_{n+1} = (a * a)_n$, so that they coincide with the Catalan numbers C_n and we have that

$$C_n = \frac{2^n (2n-1)!!}{(n+1)!}.$$

Using the identities $(2n-1)!! \cdot (2n)!! = (2n)!$ and $(2n)!! = 2^n n!$ yields the more standard expression for the nth Catalan number

$$C_n = \frac{1}{n+1} \binom{2n}{n}.$$

Exercise 59. For each $n \geq 1$, let A_n be the number of length-n strings using the symbols "(", ")", and "$*$" such that every left parenthesis "(" is correctly matched to a right parenthesis ")". For example,

$$A_1 = \#\{*\} = 1, \qquad A_2 = \#\{**, ()\} = 2, \qquad A_3 = \#\{***, *(), ()*, (*)\} = 4, \qquad \dots$$

Find a recurrence satisfied by the sequence $(A_n)_{n \geq 0}$ and use it to find a formula for A_n.

Example 7.3. The Fibonacci sequence is defined by $a_0 = a_1 = 1$ and $a_{n+2} = a_n + a_{n+1}$ for every $n \geq 0$. Taking $f(t) = \sum_{n=0}^{\infty} a_n t^n$ to be the formal ordinary generating function of this

sequence, we have the equality of formal power series

$$\sum_{n=0}^{\infty} a_{n+2}t^n = \sum_{n=0}^{\infty} a_n t^n + \sum_{n=0}^{\infty} a_{n+1}t^n,$$

which can be rewritten using the "shifting indices" rule as

$$\frac{f-1-t}{t^2} = f + \frac{f-1}{t}. \tag{7.1}$$

Solving for f yields that

$$f = \frac{1}{1-t-t^2}.$$

Since this function is analytic at 0, has $f(0) = f'(0) = 1$, and satisfies the relation (7.1), the coefficients of its Taylor series at zero must be the Fibonacci sequence. The easy way to express f as a power series around zero is not directly via its Taylor series, but rather using partial fractions:

$$f = \frac{1}{1-t-t^2} = \frac{1/\sqrt{5}}{\frac{1+\sqrt{5}}{2} - t} + \frac{1/\sqrt{5}}{t - \frac{1-\sqrt{5}}{2}} = \frac{1}{\sqrt{5}} \sum_{n=0}^{\infty} \left(\frac{1+\sqrt{5}}{2}\right)^{n+1} t^n - \frac{1}{\sqrt{5}} \sum_{n=0}^{\infty} \left(\frac{1-\sqrt{5}}{2}\right)^{n+1} t^n,$$

so that

$$F_n = \frac{1}{\sqrt{5}} \left[\left(\frac{1+\sqrt{5}}{2}\right)^{n+1} - \left(\frac{1-\sqrt{5}}{2}\right)^{n+1} \right]$$

for every $n \geq 0$. Later on we will see a different (easier) way of doing the same computation.

7.2 Exponential generating functions

Given a sequence of real numbers $(a_n)_{n\geq 0}$, the **formal exponential generating function** of $(a_n)_{n\geq 0}$ is the formal power series $\sum_{n=0}^{\infty} \frac{a_n}{n!} t^n$. The **exponential generating function** of $(a_n)_{n\geq 0}$ is the function defined by summing this formal power series within its radius of convergence (taking the generating function to have empty domain when the radius of convergence is zero). In other words, the exponential generating function of $(a_n)_{n\geq 0}$ is the ordinary generating function of $(a_n/n!)_{n\geq 0}$.

As with ordinary generating functions, there are various identities relating operations on sequences and operations on exponential generating functions.

1. (Shifting index $\longrightarrow$ differentiation.) If $f(t) = \sum_{n=0}^{\infty} \frac{a_n}{n!} t^n$ is the formal exponential generating function of $(a_n)_{n\geq 1}$ and $k \geq 1$ then the formal kth derivative $f^{(k)}$ is the formal exponential generating function of $(a_{n+k})_{n\geq 0}$.

2. (Multiplication by n $\longrightarrow$ multiplication by $t\frac{d}{dt}$.) If $f(t)$ is the formal exponential generating function of $(a_n)_{n\geq 0}$ then $tf'(t)$ is the formal exponential generating function of

$(na_n)_{n\geq 0}$. More generally, if P is a real polynomial in one variable then the exponential generating function of $(P(n)a_n)_{n\geq 0}$ is given by $P(t\frac{d}{dt})f(t)$. (Yes – this rule is exactly the same for ordinary and exponential generating functions.)

3. (Binomial convolution $\longrightarrow$ multiplication.) If we multiply together two formal exponential generating functions $f = \sum_{n=0}^{\infty} \frac{a_n}{n!}t^n$ and $g = \sum_{n=0}^{\infty} \frac{b_n}{n!}t^n$, what does this mean in terms of a sequence operation? The formal product $f(t)g(t)$ can be written

$$f(t)g(t) = \sum_{n=0}^{\infty} \left(\sum_{k=0}^{n} \frac{a_k b_{n-k}}{k!(n-k)!} \right) t^n = \sum_{n=0}^{\infty} \left(\frac{1}{n!} \sum_{k=0}^{n} \binom{n}{k} a_k b_{n-k} \right) t^n,$$

so that $f(t)g(t)$ is the formal exponential generating function of the **binomial convolution** $a \otimes b$ of $a = (a_n)_{n\geq 0}$ and $b = (b_n)_{n\geq 0}$ defined by

$$(a \otimes b)_n = \sum_{k=0}^{n} \binom{n}{k} a_k b_{n-k}.$$

(This notation is not standard; I made it up.)

Let us see how we can use exponential generating functions to solve some counting problems. Note that it might not always be obvious whether to use ordinary or exponential generating functions to solve a given problem, although the presence of binomial convolutions is a clear sign that exponential generating functions should be appropriate.

Example 7.4.

1. The constant sequence $a_n \equiv 1$ has exponential generating function e^t.
2. The exponential sequence $a_n = \lambda^n$ has exponential generating function $e^{\lambda t}$.
3. The sequence $a_n = n!$ has exponential generating function $1/(1-t)$.

Example 7.5 (Derangements)**.** A **derangement** of $\{1, \ldots, n\}$ is a bijection $\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$ such that $\sigma(i) \neq i$ for every $0 \leq i \leq n$. (In other words, a derangement is a permutation with no fixed points.) Let D_n be the number of derangements of $\{1, \ldots, n\}$, where we set $D_0 = 1$. Since there are $n!$ bijections from $\{1, \ldots, n\}$ to itself, and we can uniquely specify any such bijection by first choosing its fixed points and then choosing a derangement of the non-fixed points, we have that

$$n! = \sum_{k=0}^{n} \binom{n}{k} D_{n-k} = (1 \otimes D)_n$$

for every $n \geq 0$. Taking exponential power series of both sides, using that $n!$ has formal exponential power series $1/(1-t)$ and 1 has formal exponential power series e^t, we obtain

that

$$e^t \sum_{n=0}^{\infty} \frac{D_n}{n!} t^n = \frac{1}{1-t}$$

and hence that

$$\sum_{n=0}^{\infty} \frac{D_n}{n!} t^n = \frac{e^{-t}}{1-t}.$$

Since e^{-t} is the formal exponential power series of $(-1)^n$ it follows by a second application of the binomial convolution rule that

$$\sum_{n=0}^{\infty} \frac{D_n}{n!} t^n = \sum_{n=0}^{\infty} \frac{1}{n!} \left(\sum_{k=0}^{n} \binom{n}{k} (-1)^{n-k} k! \right) t^n$$

and hence that

$$D_n = \sum_{k=0}^{n} \binom{n}{k} (-1)^{n-k} k! = n! \cdot \left[1 - \frac{1}{1!} + \frac{1}{2!} - \cdots + \frac{(-1)^n}{n!} \right].$$

In particular, for large n, D_n is approximately equal to $n!/e$. This means that the probability that a uniformly random permutation of $\{1, \ldots, n\}$ has no fixed points converges to $1/e$ as $n \to \infty$. In fact, noting that the error satisfies

$$\left| D_n - \frac{n!}{e} \right| = \left| \sum_{k=n+1}^{\infty} (-1)^k \frac{n!}{k!} \right| \leq \frac{1}{n+1},$$

which is less than $1/2$ for $n \geq 2$ (the last inequality follows since if $(a_n)_{n \geq 0}$ is any decreasing sequence then $|\sum_{n=0}^{\infty} (-1)^n a_n| \leq a_0$), D_n must actually be equal to the closest integer to $n!/e$ for every $n \geq 2$, and in fact this is true for $n = 1$ also.

Example 7.6 (Binary rooted trees with increasing labels)**.** Let A_n be the number of functions φ from $\{1, \ldots, n\}$ to itself such that $\varphi(1) = 1$, $\varphi(k) < k$ for every $k > 0$ and $\varphi^{-1}(k) \setminus \{1\} = \{2 \leq m \leq n : \varphi(m) = k\}$ has at most two elements for every $2 \leq k \leq n$. See Figure 7 for a visualization of these functions as flow charts, where we see that $A_1 = A_2 = 1$, $A_3 = 2$, and $A_4 = 5$. We also set $A_0 = 1$ to make the rest of the calculation work out nicely. Since the only feature of the numbers $\{1, \ldots, n\}$ that we use to define A_n is their relative order, A_n also counts the number of functions φ from *any* set Ω of n integers with minimal element n_0 to itself such that $\varphi(k) < k$ for every $k \in \Omega \setminus \{n_0\}$ and with $|\varphi^{-1}(k) \setminus \{n_0\}| \leq 2$ for every $k \in \Omega$. We call such a function an **admissible** function on Ω.

To get a recursion for A_n, we first note that if $\varphi : \{1, \ldots, n+1\}$ is admissible with

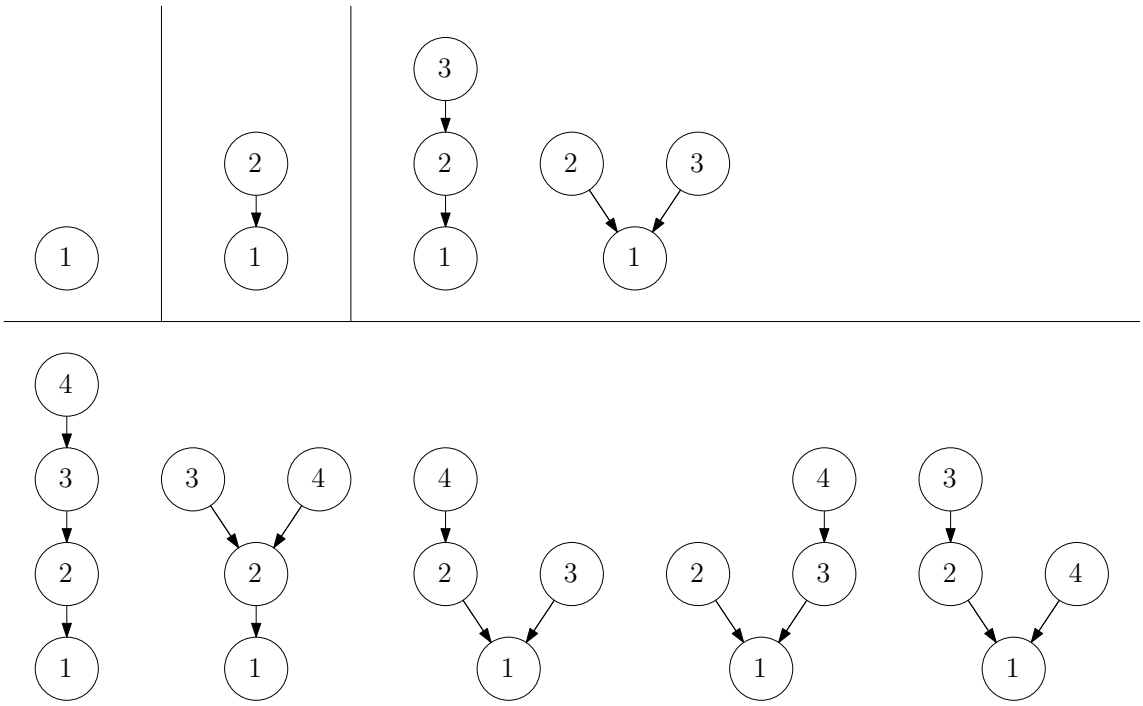


Figure 7: The functions counted in example 7.6 represented as binary rooted trees with increasing labels. Our calculation shows that the higher derivatives of tan at $\pi/4$ are secretly counting these diagrams.

$|\phi^{-1}(1) \setminus \{1\}| = 1$ then we must have that $\varphi^{-1}(1) = \{1, 2\}$ and that

$$\psi(k) = \begin{cases} \phi(k) & k \neq 2 \\ 2 & k = 2 \end{cases}$$

defines an admissible function on $\{2, \ldots, n+1\}$. This defines a bijection between admissible functions on $\{1, \ldots, n+1\}$ with $|\phi^{-1}(1) \setminus \{1\}| = 1$ and admissible functions on $\{2, \ldots, n\}$, so that the total number of such functions is A_n.

On the other hand, if $|\varphi^{-1}(1) \setminus \{1\}| = 2$, we can partition the numbers $\{2, \ldots, n+1\}$ into two sets X and Y according to the last number taken by the iterates $(k, \varphi(k), \varphi^2(k), \ldots)$ before it fixates at 1. (In other words, in the pictorial representation of Figure 7, we separate the numbers $\{2, \ldots, n+1\}$ into those appearing on the two sides of the tree.) Similarly to above, in this case the function φ can be uniquely specified by specifying the unordered pair of sets $\{X, Y\}$ whose union is $\{2, \ldots, n+1\}$, an admissible function on X and an admissible function on Y. As such, the number of admissible functions on $\{1, \ldots, n+1\}$ such that $|\varphi^{-1}(1) \setminus \{1\}| = 2$ is given by $\frac{1}{2} \sum_{k=1}^{n-1} \binom{n}{k} A_k A_{n-k}$, where the factor $1/2$ comes from the fact that the order of the pair $\{X, Y\}$ does not matter.

Putting these two identities together yields that

$$A_{n+1} = A_n + \frac{1}{2}\sum_{k=1}^{n-1}\binom{n}{k}A_kA_{n-k} = \frac{1}{2}\sum_{k=0}^{n}\binom{n}{k}A_kA_{n-k} + \frac{1}{2}\mathbb{1}(n=0),$$

where the second equality comes from the expression for the sum over the "missing terms"

$$\frac{1}{2}\sum_{k\in\{0,n\}}\binom{n}{k}A_kA_{n-k} = \begin{cases} \frac{1}{2}(A_0A_n + A_nA_0) = A_n & n > 0 \\ \frac{1}{2}A_0^2 = \frac{1}{2} & n = 0. \end{cases}$$

(Be careful: The fact that something special happens at zero [or other small values of n] is extremely easy to miss when doing these calculations!!) Thus, if $f = \sum_{n=0}^{\infty}\frac{A_n}{n!}t^n$ is the exponential generating function of $(A_n)_{n\geq 0}$ then f is a formal solution to the polynomial ODE

$$f' = \frac{1}{2}(f^2+1).$$

Being a formal solution to a polynomial ODE of the form $f' = P(f)$, the formal power series f must have positive radius of convergence, and therefore defines a real analytic function on some open interval containing 0 that solves the ODE $2f' = f^2 + 1$. Since this ODE is separable, we can solve it as a separable ODE:

$$\int \frac{2df}{f^2+1} = \int 1\,\mathrm{d}t$$

We can recognize the integral on the left as $2\tan^{-1}(f)$ and rearrange to obtain solutions of the form

$$f = \tan\left(\frac{t}{2} + C\right).$$

Since we want the solution to have $f(0) = A_0 = 1$ we must have that $\tan(C) = 1$; taking $C = \pi/4$ works. Since function and formal operations coincide within the radius of convergence and since this f is the unique solution to the ODE with $f(0) = 1$ by Picard-Lindelöf, we must have that

$$\sum_{n=0}^{\infty}\frac{A_n}{n!}t^n = \tan\left(\frac{t}{2} + \frac{\pi}{4}\right).$$

Unfortunately there is no nice expression for the Taylor series of tan, so that we do not get a nice formula for A_n. (Again, "a nice formula" does not have a precise mathematical meaning here. The double factorial $n!!$ would not look that nice either if we didn't have notation for it and just wrote it as a product.)

In the next section we will learn about what we can get out of having a nice expression for the ordinary/exponential generating function of a sequence even if this does not lead to such a nice formula; this is very similar to the question of what we can learn from the Laplace

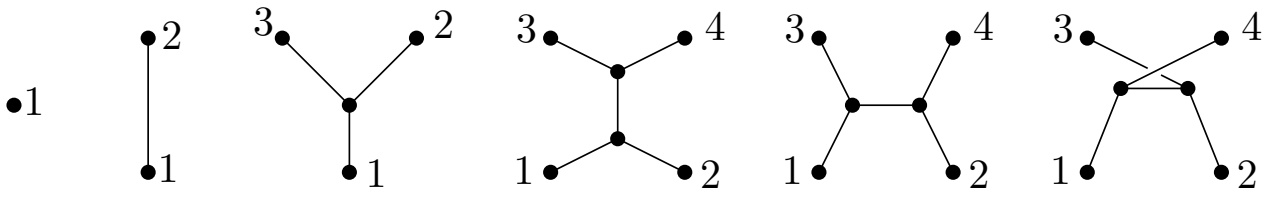
transform of a function when we are not able to explicitly invert the Laplace transform.

Exercise 60. Use exponential generating functions to solve the recursion

$$A_{n+1} = -\frac{1}{2}\sum_{k=1}^{n-1}\binom{n}{k}A_k A_{n-k+1}$$

with $A_0 = 1$.

Exercise 61. Using exponential generating functions, find a formula for the number of (isomorphism classes of) trees with n leaves labelled $1, \ldots, n$ and with unlabelled internal vertices each having degree exactly 3.



(The first few numbers in this sequence are $1, 1, 1, 3, \ldots$)

7.3 Difference equations

Given a sequence $(a_n)_{n\geq 0}$, the sequence of differences $(\Delta a_n)_{n\geq 0}$ is defined by $\Delta a_n = a_{n+1} - a_n$, so that Δ is a discrete analogue of differentiation. A **kth order difference equation** is a recurrence relation of the form

$$\Delta^k a = F(t, a, \Delta a, \ldots, \Delta^{k-1} a).$$

In practice "difference equation" is used less precisely than this, and is often just used interchangeably with "recurrence relation." The theory of difference equations is closely analogous to that of ODEs, and various simple difference equations become ODEs after we take generating functions. As in the ODE case, kth order linear difference equations can always be thought of as a first order difference in k dimensions:

$$\Delta\begin{pmatrix}\Delta^{k-1}a\\ \Delta^{k-2}a\\ \vdots\\ \Delta a\\ a\end{pmatrix} = \Delta\begin{pmatrix}F(t,a,\Delta a,\ldots,\Delta^{k-1}a)\\ \Delta^{k-1}a\\ \vdots\\ \Delta^2 a\\ \Delta a\end{pmatrix}$$

A kth order **linear** difference equation is a difference equation of the form

$$a_{n+k} + c_{k-1}(n)a_{n+k-1} + \cdots + c_0(n)a_n = b(n)$$

where $c_0, \ldots, c_{k-1}$ and b are function from $\{0, 1, \ldots\}$ to $\mathbb{R}$. (We could write this relation in terms of the differences, but there's not much reason to do so.) As in the ODE case, linear ODEs can be written in terms of matrices as

$$\begin{pmatrix} a_{n+k} \\ a_{n+k-1} \\ \vdots \\ a_{n+1} \\ 1 \end{pmatrix} = \begin{pmatrix} -c_{k-1}(n) & -c_{k-2}(n) & \cdots & -c_1(n) & -c_0(n) & b(n) \\ 1 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & & & & \\ 0 & 0 & \cdots & 1 & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a_{n+k-1} \\ a_{n+k-2} \\ \vdots \\ a_n \\ 1 \end{pmatrix},$$

so that if we write $\mathbf{a}_n = (a_{n+k-1}, \cdots, a_n)$ and write $A(n)$ for the matrix appearing here then

$$\begin{pmatrix} \mathbf{a_{n+1}} \\ 1 \end{pmatrix} = A(n) \begin{pmatrix} \mathbf{a_n} \\ 1 \end{pmatrix} = A(n)A(n-1)\cdots A(0) \begin{pmatrix} \mathbf{a_0} \\ 1 \end{pmatrix}.$$

If $A(n) = A$ is constant, then this simplifies to

$$\begin{pmatrix} \mathbf{a_{n+1}} \\ 1 \end{pmatrix} = A^n \begin{pmatrix} \mathbf{a_0} \\ 1 \end{pmatrix}.$$

So that solving constant-coefficient linear difference equations reduces to taking *powers* of matrices just as solving constant-coefficient linear ODEs reduces to taking *exponentials* of matrices; such powers can be computed efficiently by diagonalizing / taking the Jordan normal form as appropriate just as we did when exponentiating.

Exercise 62. Use this method to derive the formula for the nth Fibonacci number.

Note also that if a solves a constant coefficient linear difference equation

$$a_{n+k} + c_{k-1}a_{n+k-1} + \cdots + c_0 a_n = b$$

then the formal exponential generating function $f(t) = \sum_{n=0} \frac{a_n}{n!} t^n$ of a satisfies the constant coefficient linear ODE

$$f^{(k)} + c_{k-1}f^{(k-1)} + \cdots + c_0 f = be^t.$$

This means that there is a sense in which constant coefficient linear ODEs and difference equations are *equivalent*, rather than merely analogous.

8 Introduction to asymptotic analysis

The goal of this section is to provide a brief introduction to *asymptotic analysis* and *Tauberian theory*. Asymptotic analysis is a huge topic that could easily be the subject of an entire course, so we will only be scratching the surface.

8.1 Asymptotic notation

We begin with a primer on asymptotic notation. Asymptotic notation is extremely useful throughout analytic branches of mathematics, and is what modern researchers in probability, combinatorics, PDE, theoretical computer science theory, analytic number theory, etc. use to describe their work. Part of the utility of this notation is that it is very flexible, which makes it difficult to provide a single unified formal framework for its use, and you should not always expect other uses of this notation that you see to strictly conform to the rules we set out here.

Here are the basics. Suppose that we have **non-negative** functions f and g defined on domains including a neighbourhood of some point $t \in [-\infty, +\infty]$; if $t = +\infty$ or $-\infty$ this means that their domains include all sufficiently large positive or negative numbers as appropriate. "Big-O", "little-o", "big-Ω", and "little-ω" notation are defined as follows:

$$\begin{array}{lll}
\text{“}f(t) = O(g(t)) \text{ as } t \to t_0\text{”} & \text{means that} & \limsup_{t \to t_0} \frac{f(t)}{g(t)} < \infty \\
\text{“}f(t) = o(g(t)) \text{ as } t \to t_0\text{”} & \text{means that} & \limsup_{t \to t_0} \frac{f(t)}{g(t)} = 0 \\
\text{“}f(t) = \Omega(g(t)) \text{ as } t \to t_0\text{”} & \text{means that} & \liminf_{t \to t_0} \frac{f(t)}{g(t)} > 0 \\
\text{“}f(t) = \omega(g(t)) \text{ as } t \to t_0\text{”} & \text{means that} & \liminf_{t \to t_0} \frac{f(t)}{g(t)} = \infty.
\end{array}$$

We also define "Θ" notation (there is no little-θ)

$$\text{“}f(t) = \Theta(g(t)) \text{ as } t \to t_0\text{”} \text{ means that } f(t) = O(g(t)) \text{ and } f(t) = \Omega(g(t)) \text{ as } t \to t_0.$$

Finally, we write

$$\text{“}f(t) \sim g(t) \text{ as } t \to t_0\text{”} \text{ to mean that } \lim_{t \to t_0} \frac{f(t)}{g(t)} = 1,$$

in which case we say f is asymptotically equal to g as $t \to t_0$. All these notions extend in obvious ways to situations where e.g. one has sequences instead of continuous-domain functions, or where one requires t to converge to t_0 from a particular direction. Often one omits the "as $t \to t_0$" part if the relevant limit is obvious from context.

Some warnings are in order:

1. The requirement that quantities written in this notation are always non-negative is not completely standard, but greatly increases the expressive power of the notation. If we want to talk about quantities of uncertain sign we can use $\pm$, so that e.g.

$$f(x) = \pm O(x^2) \text{ as } x \to +\infty$$

means that $|f(x)| = O(x^2)$ as $x \to +\infty$.

2. In some fields one often sees big-O notation used to mean what we denote Θ here (often in the same places where it is also used in the same way we use big-O). We much prefer having two different notations for the two different things.

Another feature of this notation is that we can use expressions like $O(g(t))$ etc. to stand in for a function satisfying the appropriate asymptotic bound as part of a more complicated expression. So, for example,

$$f(t) = f(t_0) + (t - t_0)f'(t_0) \pm o(|t - t_0|) \qquad \text{as } t \to t_0$$

means that there exists a function $h(t)$ with $|h(t)| = o(|t - t_0|)$ as $t \to t_0$ such that $f(t) = f(t_0) + (t - t_0)f'(t_0) + h(t)$, and

$$a_n = e^{(1+o(1))n} \qquad \text{as } n \to \infty$$

means that there exists a non-negative sequence $h(n)$ with $h(n) \to 0$ as $n \to \infty$ such that $a_n = e^{(1+h(n))n}$. Moreover, $f(t) \sim g(t)$ as $t \to t_0$ if and only if $f(t) = (1 \pm o(1))g(t)$ as $t \to t_0$.

8.2 Asymptotic expansions

As we saw with the example $f(x) = e^{-x^{-2}}$, it is not always true that a smooth function can always be recovered from its Taylor series around a point (not every smooth function is real analytic). Still, Taylor's theorem does tell us that if f is smooth at x_0 then

$$f(x) = \sum_{n=0}^{N} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n \pm o(|x - x_0|^N)$$

as $x \to x_0$ for each $N \geq 0$. Thus, there is still some weak sense in which f is "equal" to the infinite sum $\sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$, even when this sum is not well-defined! This motivates the definition of an *asymptotic expansion*, of which the Taylor series is a key example.

Just as we previously considered formal power series, we can now consider arbitrary formal series of functions $\sum_{n=0}^{\infty} a_n f_n$. (Rigorously, this is 'just notation' for a sequence of functions $(f_0, f_1, \ldots)$ and coefficients $(a_0, a_1, \ldots)$. We will put more conditions on these functions if we want to define multiplication and differentiation formally.) Let $t_0 \in [-\infty, \infty]$ and let $\Omega \subseteq \mathbb{R}$ be an open interval which either contains t_0 or has t_0 as an endpoint. Given a function f and

a sequence of functions $(f_n)_{n\geq 0}$ defined on Ω such that $|f_{n+1}(t)| = o(|f_n(t)|)$ as $t \to t_0$ and a sequence of real numbers $(a_n)_{n\geq 0}$, we say that

$$f(t) = \sum_{n=0}^{\infty} a_n f_n(t) \mod \text{t.s.t.s} \qquad \text{asymptotically as } t \to t_0$$

if

$$f(t) = \sum_{n=0}^{N} a_n f_n(t) \pm o(|f_N(t)|) \qquad \text{as } t \to t_0 \text{ for every } N \geq 0$$

or equivalently if

$$f(t) = \sum_{n=0}^{N} a_n f_n(t) \pm O(|f_{N+1}(t)|) \qquad \text{as } t \to t_0 \text{ for every } N \geq 0.$$

Exercise 63. Prove that these two definitions are equivalent.

The "equality" $f(t) = \sum_{n=0}^{\infty} f_n(t)$ is referred to as an **asymptotic expansion** of $f(t)$. The "t.s.t.s" in this notation stands for "transcendentally small terms", and refers to the fact that if f and g are two functions satisfying $|f(t) - g(t)| = o(|f_k(t)|)$ as $t \to t_0$ for every $k \geq 0$ then we can have that both

$$f(t) = \sum_{n=0}^{\infty} a_n f_n(t) \mod \text{t.s.t.s} \qquad \text{asymptotically as } t \to t_0$$

and

$$g(t) = \sum_{n=0}^{\infty} a_n f_n(t) \mod \text{t.s.t.s} \qquad \text{asymptotically as } t \to t_0$$

even when f and g are not actually equal. Be careful to note that what constitutes a transcendentally small term depends on the sequence of functions $(f_n)_{n\geq 0}$. As before, all these notions have alternative versions making sense for sequences and for one-sided convergence $t \uparrow t_0$ or $t \downarrow t_0$.

Remark 8.1. The standard notation for asymptotic expansions is just

$$f(t) \sim \sum_{n=0}^{\infty} a_n f_n(t) \qquad \text{as } t \to t_0.$$

Our (highly non-standard) notation is chosen to avoid overloading the $\sim$ notation (since, in our usage, the simpler relation $f(t) \sim a_0 f_0(t)$ holds in this case when $a_0 \neq 0$) and to stress that anything going to zero faster than everything used in the sequence of functions used in the expansion is invisible to the expansion.

Note that if

$$f(t) = \sum_{n=0}^{\infty} a_n f_n(t) \mod \text{t.s.t.s} \qquad \text{asymptotically as } t \to t_0$$

for some sequence of coefficients $(a_n)_{n\geq 0}$ then we can compute these coefficients from f by

$$a_0 = \lim_{t\to t_0} \frac{f(t)}{f_0(t)}, \qquad a_1 = \lim_{t\to t_0} \frac{f(t) - a_0 f_0(t)}{f_1(t)}, \qquad a_1 = \lim_{t\to t_0} \frac{f(t) - a_0 f_0(t) - a_1 f_1(t)}{f_2(t)}, \quad \dots$$

In particular, these coefficients are unique (given f and $f_0, f_1, \dots$) when they exist.

Example 8.2. Taylor's Theorem implies that if f is smooth at a point $t_0 \in \mathbb{R}$ then

$$f(t) = \sum_{n=0}^{\infty} \frac{f^{(n)}(t_0)}{n!}(t-t_0)^n \mod \text{t.s.t.s} \qquad \text{asymptotically as } t \to t_0.$$

Note that this makes sense even if the Taylor series has zero radius of convergence!

Exercise 64. Prove that if $f(t) = \sum_{n=0}^{\infty} a_n f_n(t)$ mod t.s.t.s as $t \to t_0$ and $f(t) = \sum_{n=0}^{\infty} b_n f_n(t)$ mod t.s.t.s as $t \to t_0$ then $a_n = b_n$ for every $n \geq 0$. (Recall the standing assumption that $|f_{n+1}(t)| = o(|f_n(t)|)$ as $t \to t_0$ for each $n \geq 0$.)

Example 8.3 (The exponential integral)**.** Let us try to come up with an asymptotic expansion for the function

$$f(t) = \int_1^t \frac{e^s}{s}\,\mathrm{d}s.$$

The first term in this expansion is fairly straightforward to see: Since $\frac{e^t}{t}$ increases very rapidly, most the contribution to the integral will come from values of s that are very close to t, so that we should have

$$\int_1^t \frac{e^s}{s}\,\mathrm{d}s \sim \frac{1}{t}\int_1^t e^s\,\mathrm{d}s = \frac{1}{t}(e^t - 1) \sim \frac{e^t}{t}.$$

To prove this rigorously, we can fix $\varepsilon > 0$ and write

$$\begin{aligned}
\left|\int_1^t \frac{e^s}{s}\,\mathrm{d}s - \frac{1}{t}\int_1^t e^s\,\mathrm{d}s\right| &\leq \int_1^t \left|\frac{1}{s} - \frac{1}{t}\right| e^s\,\mathrm{d}s \\
&\leq \int_{t/(1+\varepsilon)}^t \left|\frac{1}{s} - \frac{1}{t}\right| e^s\,\mathrm{d}s + \int_1^{t/(1+\varepsilon)} \left|\frac{1}{s} - \frac{1}{t}\right| e^s\,\mathrm{d}s \\
&\leq \int_{t/(1+\varepsilon)}^t \left|\frac{\varepsilon}{t}\right| e^s\,\mathrm{d}s + \int_1^{t/(1+\varepsilon)} e^s\,\mathrm{d}s \\
&\leq \varepsilon \int_1^t \frac{e^s}{t}\,\mathrm{d}s + e^{t/(1+\varepsilon)}. \qquad (8.1)
\end{aligned}$$

Since $\varepsilon > 0$ was arbitrary and $e^{t/(1+\varepsilon)} = o(e^t)$ for each fixed $\varepsilon > 0$, we can deduce that

$$\left| \int_1^t \frac{e^s}{s}\,\mathrm{d}s - \frac{1}{t}\int_1^t e^s\,\mathrm{d}s \right| = o\left(\frac{e^t}{t}\right) \tag{8.2}$$

and hence that

$$\int_1^t \frac{e^s}{s}\,\mathrm{d}s \sim \frac{1}{t}\int_1^t e^s\,\mathrm{d}s \sim \frac{e^t}{t}.$$

as required. Since this kind of thinking is new to you, it will probably not be obvious to you how to get from (8.1) to (8.2): The point is that if we take $\varepsilon(t)$ to go to zero sufficiently slowly as $t \to \infty$ (in this example taking $\varepsilon = c\log t/t$ for a sufficiently small constant t works) then the right hand side will be $o(e^t/t)$ as required.

In fact for this example there is a nice way to compute the asymptotics in a much less messy way that gives the entire asymptotic series: Integration by parts! We have that

$$\int_1^t \frac{e^s}{s}\,\mathrm{d}s = \left[\frac{1}{t}e^t\right]_1^t + \int_1^t \frac{e^s}{s^2}\,\mathrm{d}s,$$

and iterating this calculation any finite number of times

$$\int_1^t \frac{e^s}{s}\,\mathrm{d}s = \sum_{n=1}^{N}\left[\frac{(n-1)!}{t^n}e^t\right]_1^t + \int_1^t \frac{N!}{s^{N+1}}e^s\,\mathrm{d}s,$$

for every $N \geq 1$. This means that $\int_1^t \frac{e^s}{s}\,\mathrm{d}s$ has the (divergent) asymptotic expansion

$$\int_1^t \frac{e^s}{s}\,\mathrm{d}t = \sum_{n=0}^{\infty}\frac{n!}{t^{n+1}}e^t \quad \text{mod t.s.t.s} \qquad \text{asymptotically as } t \to +\infty.$$

Note that e.g. $\int_2^t \frac{e^s}{s}\,\mathrm{d}s$ differs from $\int_1^t \frac{e^s}{s}\,\mathrm{d}s$ by the constant $\int_1^2 \frac{e^s}{s}\,\mathrm{d}s$ and has the same asymptotic expansion.

Let us compare this to the (convergent) expansion

$$\int_1^t \frac{e^s}{s}\,\mathrm{d}s = \int_1^t \sum_{n=0}^{\infty}\frac{s^{n-1}}{n!}\,\mathrm{d}s = \log t + \sum_{n=1}^{\infty}\frac{t^n - 1}{n\cdot n!}$$

Although this expansion converges, it is *not* an asymptotic expansion as $t \to +\infty$: If we want to get an accurate estimate of our function using this expansion we need to take a large number of terms when t is large.

Exercise 65. Prove that

$$\sum_{n=0}^{\lfloor(1+\varepsilon)t\rfloor}\frac{t^n}{n!} \sim e^t \qquad \text{as } t \to +\infty \qquad \text{and} \qquad \sum_{n=0}^{\lfloor(1-\varepsilon)t\rfloor}\frac{t^n}{n!} = o(e^t) \qquad \text{as } t \to +\infty$$

for every $\varepsilon > 0$, where $\lfloor x \rfloor$ means x rounded down to the nearest integer.

Remark 8.4. In fact we have more precisely that

$$\sum_{n=\lfloor t-m(t) \rfloor}^{\lceil t+m(t) \rceil} \frac{t^n}{n!} \sim e^t \text{ as } t \to +\infty \qquad \text{if and only if} \qquad m(t) = \omega(t^{1/2}) \text{ as } t \to +\infty.$$

(This can be seen as a consequence of the central limit theorem!) This means that we need to use about $\sqrt{t}$ terms to get a good estimate to e^t from its series expansion when t is large.

Example 8.5. You are probably familiar with Stirling's formula

$$n! \sim \sqrt{2\pi n} \left(\frac{n}{e}\right)^n.$$

This asymptotic formula can be expressed in terms of $\log(n!)$ as

$$\log(n!) = n \log n - en + \frac{1}{2} \log n + \frac{1}{2} \log 2\pi \pm o(1).$$

(In general, if we want to determine the first-order asymptotics of some quantity, we should expand its logarithm up to $\pm o(1)$ terms.) There is in fact an asymptotic expansion

$$\log n! = n \log n - en + \frac{1}{2} \log n + \frac{1}{2} \log 2\pi + \sum_{k=2}^{\infty} \frac{(-1)^k B_k}{k(k-1)n^{k-1}} \quad \text{mod t.s.t.s} \qquad \text{asymptotically as } n \to \infty,$$

where B_k is the kth *Bernoulli number*. This is a divergent asymptotic series expansion.

8.3 Formal series solutions beyond power series

Up until now we have discussed formal solutions to ODEs only in terms of formal *power series* solutions $\sum_{n=0}^{\infty} a_n t^n$ or $\sum_{n=0}^{\infty} a_n (t-t_0)^n$. While power series solutions are very flexible, they do not always apply as we saw for the ODE $tf' - 2f - 2t^2$, which did not have any formal solutions but did have the solution $t^2 \log t^2$, which is not real analytic at zero. Moreover, while power series are obviously useful when studying the behaviour of our function near zero, they might not be the most appropriate way of understanding the *large time* behaviour of our function, where other expansions might be more appropriate.

For all these reasons, it is often natural to work with other kinds of series solutions to ODEs besides power series. To do this with *formal* series, one must of course work within a class of formal series for which all the operations needed to define the ODE (most commonly differentiation and multiplication) are well-defined. For example one can work with formal

power series $\sum_{n=0}^{\infty} a_n t^{-n}$ in t^{-1}, for which the theory is much the same as for formal power series as we studied them before.

To identify the first term in our series solution, we can attempt to find something which is "almost" a solution to the ODE, then recursively attempt to "correct" this solution into an actual solution by adding in more terms.

Regular singular points and the Frobenius method. Consider an nth order homogeneous linear ODE of the form

$$P_n(t)f^{(n)} + P_{n-1}(t)f^{(n-1)} + \cdots + P_0(t)f = 0,$$

where $P_n, \ldots, P_0$ are polynomials[14]. A simple and important example is

$$tf' - \alpha f = 0$$

for a constant $\alpha \in \mathbb{R}$, which has solutions of the form Ct^α that do not have a power series representation at zero when α is not an integer and $C \neq 0$. (In particular, for non-integer α this equation does have a single formal power series solution – the all zero power series – but these solutions do not recover all solutions to the ODE.)

When $P_n(t) \neq 0$, we can write the ODE in the form

$$f^{(n)} + \frac{P_{n-1}}{P_n}f^{(n-1)} + \cdots + \frac{P_0}{P_n}f = f^{(n)} + Q_{n-1}f^{(n-1)} + \cdots + Q_0 f = 0 \tag{8.3}$$

where $Q_0, \ldots, Q_{n-1}$ are rational functions. We say that $t_0 \in \mathbb{R}$ is an **ordinary point** for the ODE if $Q_i(0)$ is well-defined for every $0 \leq i \leq n-1$. Note that every point which is not a zero of P_n must be a regular point, but that cancellations between factors appearing in both the numerator and denominator of each Q_i can lead to some zeros of P_n also being regular points. The ordinary points of the ODE are precisely when our usual theory of power series solutions works with only minor modifications:

Exercise 66. Prove that if 0 is an ordinary point then the ODE has an n-dimensional space of formal power series solutions.

Points that are not ordinary are called **singular points**. We would still like to find some kind of series solution for our ODE around a singular point, but know that power series will not always work. How, then, should we identify the correct sequence of functions to take our expansion with respect to? If we want our expansion to be asymptotic, the first term should describe the first-order asymptotics of our solution as $t \to 0$. One way to proceed, in light of the example above, is to guess that our solutions can be expanded in possibly non-integer powers of t as $t \to 0$. If f does have a convergent expansion of the form $f(t) = \sum_{n=0}^{\infty} a_n t^{\alpha_n}$

[14]Everything we discuss here applies equally well with polynomials replaced by analytic functions and rational functions replaced by so-called *meromorphic* functions.

for some strictly increasing sequence $\alpha_0 < \alpha_1 < \cdots$ then the first-order asymptotics of f and each of its derivatives as $t \to 0$ are determined by the first of these powers $\alpha = \alpha_0$:

$$f^{(n)}(t) = (1 \pm o(1))(\alpha - n + 1)\frac{f^{(n-1)}(t)}{t} = \cdots = (1 \pm o(1))[\alpha(\alpha-1)\cdots(\alpha-n+1)]\frac{f(t)}{t^n}$$
$$= (1 \pm o(1))(\alpha)_n \frac{f(t)}{t^n}$$

as $t \to 0$ for each $n \geq 1$, where $(\alpha)_n := \alpha(\alpha-1)\cdots(\alpha-n+1)$ denotes the **falling factorial**. (When α is an integer we might not be able to write these relations using the notation $\sim$ as $f \sim 0 \cdot g$ is not equivalent to $f = o(1)g$.) Plugging this into the ODE yields that

$$(1 \pm o(1))(\alpha)_n \frac{f}{t^n} + (1 \pm o(1))(\alpha)_{n-1}\frac{P_{n-1}}{P_n}\frac{f}{t^{n-1}} + \cdots + (1 \pm o(1))\frac{P_0}{P_n} f. \tag{8.4}$$

Multiplying both sides by t^n/f, we have that

$$(1 \pm o(1))(\alpha)_n + (1 \pm o(1))(\alpha)_{n-1}\frac{P_{n-1}}{P_n}t + \cdots + (1 \pm o(1))\frac{P_0}{P_n}t^n. \tag{8.5}$$

We say that 0 is a **regular singular point** if is not an ordinary point but each of the terms $\frac{P_{n-i}}{P_n}t^i$ appearing on the right hand side has a finite limit as $t \to 0$. (Equivalently, if the rational function $\frac{P_{n-i}}{P_n}$ either has no singularity at 0 or a pole of order at most i at 0 for each $1 \leq i \leq n$. This also defines what it means for a point other than zero to be a regular singular point). Singular points that are not regular singular points are called **irregular singular points**. If 0 is a regular singular point and we write $c_{n-i} = \lim_{t\to 0}\frac{P_{n-i}}{P_n}t^i$ for each $1 \leq i \leq n$, then we can take the $t \to 0$ limit in (8.5) to obtain that

$$(\alpha)_n + c_{n-1}(\alpha)_{n-1} + \cdots + \alpha c_1 + c_0 = 0. \tag{8.6}$$

This is known as the **indicial equation**; the calculations we've done suggest that if the ODE has a series solution in (possibly non-integer) powers of t then the first power in this series should be a solution to the inidicial equation. (Note that this equation may have complex or negative roots; in the case of complex roots it does not really make sense to speak of the first root but we'll come back to this later.)

Before going further, it will be helpful to set up an appropriate class of formal series to describe rational functions that may have singularities at 0. A **(formal) Laurent polynomial** is a (formal) series of the form

$$\sum_{n=-\infty}^{\infty} a_n t^n$$

with $a_n = 0$ for all sufficiently large negative numbers n. Every rational function can be expanded as a Laurent polynomial, with the maximal $n \geq 0$ such that $a_{-n} \neq 0$ being the order of the pole at zero if there is one. We can define addition, multiplication, and differentiation

of formal Laurent polynomials exactly as we did for formal power series.

We now go back to our ODE. Suppose that we attempt to write down a solution as a series beginning with a multiple of t^α. If we start differentiating t^α we will naturally obtain terms of the form $t^{\alpha-1}$, $t^{\alpha-2}$, etc., and multiplying our functions by rational functions will lead more generally to terms of the form $t^{\alpha+n}$ for arbitrary integers n. Thus, it makes sense to try to find solutions described by formal series $\sum_{n=0}^{\infty} a_n t^{\alpha+n}$. If we consider more generally formal series of the form $\sum_{n=-\infty}^{\infty} a_n t^{\alpha+n}$ with $a_n = 0$ for all sufficiently large negative n, then we have well-defined formal operations describing differentiation, addition, and multiplication by formal power series:

$$\frac{d}{dt} \sum_{n=-\infty}^{\infty} a_n t^{\alpha-n} := \sum_{n=-\infty}^{\infty} (\alpha - n - 1) a_{n+1} t^{\alpha-n}$$

$$(\sum_{n=-\infty} a_n t^{\alpha-n})(\sum_{n=-\infty}^{\infty} b_n t^n) := \sum_{n=-\infty}^{\infty} \left(\sum_{m=-\infty}^{\infty} a_m b_{n-m} \right) t^{\alpha-n} = \sum_{n=-\infty}^{\infty} (a * b) t^{\alpha-n}$$

where the coefficient in the product is a sum over finitely many terms since our series are assumed to terminate for sufficiently large negative indices. (Note that the convolution here is being used in a slightly more general sense than previously.) (Note that if we took *products* of these series we would get a more kind of series involving powers of the form $t^{k\alpha+n}$ for integers $k, n \geq 0$. In the example we are currently considering such products do not arise because the ODEs are linear with analytic coefficients.) Note that this all makes sense even if α is complex.

If we expand each of our rational functions $Q_i = P_i/P_n$ as a Laurent polynomial

$$Q_i = \sum_{k=-(n-i)}^{\infty} q_{i,k} t^k,$$

where we have used the assumption that 0 is a regular singular point to rule out terms with negative powers larger than $n - i$, we have a well-defined notion of what it means for $\sum_{k=0}^{\infty} a_k t^{\alpha+k}$ to be a formal solution to the ODE:

$$\sum_{k=-\infty}^{\infty} \left[(\alpha + k + n)_n a_{k+n} + \sum_{\ell=-\infty}^{\infty} q_{n-1,\ell} (\alpha + k - \ell)_{n-1} a_{k-\ell+n-1} + \cdots + \sum_{\ell=-\infty}^{\infty} q_{0,\ell} a_{k-\ell} \right] t^{\alpha+k} = 0,$$

or in other words that each coefficient appearing here is zero. The coefficients $q_{i,n-i}$ are precisely the constants c_i we defined earlier. Using that 0 is a regular singular point we can

write

$$\sum_{k=-\infty}^{\infty}\left[(\alpha+k+n)_n a_{k+n}+\sum_{\ell=-1}^{\infty} q_{n-1,\ell}(\alpha+k-\ell)_{n-1}a_{k-\ell+n-1}+\cdots+\sum_{\ell=-n}^{\infty} q_{0,\ell}a_{k-\ell}\right]t^{\alpha+k}=0, \tag{8.7}$$

so that the kth coefficient depends only on a_i for $i \leq k+n$, and the coefficient of a_{k+n} in this coefficient is given by

$$(\alpha+k+n)_n + q_{n-1,1}(\alpha+k+n-1)_{n-1}+\cdots+q_{0,n}. \tag{8.8}$$

The assumption that α solves the indicial equation is equivalent to (8.7) holding for the term $k=-n$, while all coefficients for $k<-n$ are trivially zero. Thus, there is no constraint on a_0, and **if the coefficient** (8.8) **is non-zero for every** $k>-n$ then we can solve uniquely for the remaining coefficients. Note that if this quantity does vanish at some k, then $\alpha+k$ must be another solution to the indicial equation.

Proposition 8.6. *If no two roots of the indicial polynomial are separated by an integer, then for each root α of the indicial equation and each $a_0 \in \mathbb{C}$ there is a unique (possibly complex) formal solution to the ODE of the form $\sum_{k=0} a_k t^{\alpha+k}$.*

In particular, if the indicial polynomial has n distinct roots none of which are separated by an integer, then we have obtained n different one-dimensional spaces of formal solutions, so that we should be able to obtain all solutions to the ODE by summing these solutions. (In the case of complex roots, we should be able to sum complex-conjugate pairs to get real solutions.) In fact this does always work, and the series expansions we obtain this way are always convergent. Solving the equation at a regular singular point in this way is known as the **Frobenius method**.

In the case that there are roots separated by an integer, the Frobenius method will still work to produce a unique solution when we use, say, a root of maximal real part. In the second order case, it turns out that if the two roots are separated by an integer (or equal) and f is the solution corresponding to the larger of the two roots then a second solution can be given in the form

$$f\log t+\sum_{k=0}^{\infty} a_k t^{\alpha+k}$$

where α is the smaller of the two roots.

Example 8.7. Consider the second-order, linear ODE $4tf''+2f'+f=0$. Writing this equation in the form

$$f''+\frac{1}{2t}f'+\frac{1}{4t}f=0,$$

the indicial equation is

$$\alpha(\alpha-1)+\frac{1}{2}\alpha=0$$

which has roots $\alpha = 0$ and $\alpha = 1/2$. Looking first for a formal power series solution $\sum_{n=0}^{\infty} a_n t^n$, we obtain that

$$\sum_{n=-\infty}^{\infty} \left[(n+2)(n+1)a_{n+2} + \frac{1}{2}(n+2)a_{n+2} + \frac{1}{4}a_{n+1} \right] t^n = 0,$$

where we set $a_n = 0$ for $n < 0$. This has the solution

$$a_{n+2} = -\frac{1}{4(n+2)(n+1) + 2(n+2)} a_{n+1} = -\frac{1}{(2n+4)(2n+3)} a_{n+1}$$
$$= \cdots = \frac{(-1)^{n+2}}{(2n+4)!} a_0$$

for every $n \geq -1$. This yields a convergent series (with infinite radius of convergence) and hence a function solution of the ODE. In fact this solution is given by

$$a_0 \sum_{n=0}^{\infty} \frac{(-1)^{n+2}}{(2n+4)!} t^n = a_0 \sum_{n=0}^{\infty} \mathbb{1}(n \text{ even}) \frac{(-1)^{(n/2)}}{n!} (t^{1/2})^n = a_0 \cos(\sqrt{t}).$$

We next want to find a formal solution of the form $\sum_{n=0}^{\infty} b_n t^{1/2+n}$. In this case our formal solution must satisfy

$$\sum_{n=-\infty}^{\infty} \left[(1/2+n+2)(1/2+n+1)b_{n+2} + \frac{1}{2}(1/2+n+2)b_{n+2} + \frac{1}{4}b_{n+1} \right] t^{1/2+n} = 0.$$

$$b_{n+2} = -\frac{1}{(2n+5)(2n+3) + (2n+5)} b_{n+1} = \frac{1}{(2n+5)(2n+4)} b_{n+1} = \cdots = \frac{(-1)^{n+2}}{(2n+5)!} b_0$$

Again this series has infinite radius of convergence, so that it defines a function solution to the ODE, and in fact we can recognise this solution as $b_0 \sin(\sqrt{t})$. Thus, we have solutions to the ODE of the form

$$a_0 \cos(\sqrt{t}) + b_0 \sin(\sqrt{t}).$$

These can be shown to give all solutions to the ODE. (Note that Picard-Lindelof does not obviously apply.)

8.4 A non-linear example

[This section was lectured only briefly in 2024. A revised and briefer version is to be written.]

If we want to solve a particular ODE it is not always clear what kind of series we should use. Since the first term in the series should describe the first-order asymptotics of the solution, correctly identifying this first term should be is equivalent to determining these asymptotics. There are some good semi-rigorous ways of doing this, which are very useful for guessing the right form for a series solution before trying to prove that it really works. We say that a

function is **regularly varying** of index α as $t \to +\infty$ if f is eventually non-zero and

$$\lim_{t\to\infty} \frac{f(\lambda t)}{f(t)} = \lambda^\alpha$$

for each $\lambda > 0$; f is said to be regularly varying if it is regularly varying for some index $\alpha \in \mathbb{R}$. A regularly varying function of index 0 is called a **slowly varying** function, so that every regularly varying function of index α can be written $f(t) = L(t)t^\alpha$ for some slowly varying function L. For example, t^α is a regularly varying function of index α, while $\log t$ is a slowly varying function. Note that if f is regularly varying and $g(t) \sim f(t)$ as $t \to +\infty$ then g is regularly varying with the same index as f.

Exercise 67. Prove that if f is continuously differentiable, eventually non-zero, and satisfies $(\log f)' = (1 \pm o(1))\alpha/t$ as $t \to \infty$ then f is regularly varying of index α.

Exercise 68. Prove that a continuous function $f : (0, \infty) \to \mathbb{R}$ is regularly varying if and only if it is eventually non-zero and

$$\lim_{t\to+\infty} \frac{f(\lambda t)}{f(t)}$$

is well-defined for every $\lambda > 0$.

Exercise 69. Prove that if L is a continuous slowly varying function then $L(t) = t^{\pm o(1)}$ as $t \to \infty$. Is the converse true? Deduce that if f and g are regularly varying functions and the index of f is strictly smaller than the index of g then $|f(t)| = o(|g(t)|)$ as $t \to +\infty$.

Exercise 70. Let $f(t) = L(t)t^\alpha$ be a regularly varying continuous function of index α defined on $[0, \infty)$ such that $L(t) \neq 0$ for all sufficiently large t.

1. If $\alpha > -1$ then
$$\int_0^t f(s)\,\mathrm{d}s \sim \frac{L(t)t^{\alpha+1}}{\alpha+1}$$
as $t \to +\infty$.

2. If $\alpha < -1$ then $\int_0^t f(s)\,\mathrm{d}s$ converges to a constant as $t \to +\infty$ and
$$\int_t^\infty f(s)\,\mathrm{d}s \sim \frac{L(t)t^{\alpha+1}}{\alpha+1}$$
as $t \to +\infty$.

3. If $\alpha = -1$ then either $\int_0^\infty |f(s)|\,\mathrm{d}s = \infty$ and $\int_0^t f(s)\,\mathrm{d}s$ is slowly varying or $\int_0^\infty |f(s)|\,\mathrm{d}s < \infty$ and $\int_t^\infty f(s)\,\mathrm{d}s$ is slowly varying.

In particular, $\int_0^t f(s)\,\mathrm{d}s$ either converges to a non-zero constant or is regularly varying of index $\alpha + 1$.

Very often, in the absence of oscillation, the solution to an ODE is either a regularly varying function or the exponential of a regularly varying function, and the same is true for the higher derivatives of the function. If we *assume* this is the case, we can often use the above exercise to compute the asymptotics of the solution conditional on this assumption.

It's time to try this out in an innocuous-looking example: Stokes equation $f'' = tf$. This turns out to be fairly involved, so buckle up for some calculations!

Example 8.8. Let us consider Stokes' equation $f'' = tf$. A solution to this equation cannot have second derivative that is continuous and regularly varying with respect to any index: If f'' is regularly varying with index $\alpha \geq -1$ then tf is regularly varying with index $\alpha+3$, while if $\alpha < -1$ then $f'(t)$ converges to a constant but f is regularly varying with exponent $\alpha - 1$; this can occur only if $f'(t)$ converges to zero, in which case $f'(t) = -\int_t^\infty f''(s)\,\mathrm{d}s$ is regularly varying of exponent $\alpha + 1$ and f is either converging to a non-zero constant or is regularly varying of exponent $\alpha + 2$, both of which are inconsistent with the ODE $f'' = tf$ and the assumption that f'' is regularly varying with exponent α.

The next most natural thing to try is that f is the exponential of a regularly varying function, so that $f(t) = e^{g(t)}$ for some $g = \log f$. The function g must satisfy the ODE

$$(g'' + (g')^2)e^{g(t)} = te^{g(t)}$$

and cancelling the $e^{g(t)}$ from both sides yields that

$$g'' + (g')^2 = t.$$

To understand the large-time asymptotics of the solutions to this equation, we can try **assuming** that our solution satisfies $g'' = L(t)t^\alpha$ for some slowly varying function L. If $\alpha > -1$ then we have that

$$L(t)t^\alpha + \frac{(1 \pm o(1))L(t)^2 t^{2\alpha+2}}{(\alpha+1)^2} = t$$

and since $L(t) = t^{\pm o(1)}$ we can simplify this to

$$\frac{(1 \pm o(1))L(t)^2 t^{2\alpha+2}}{(\alpha+1)^2} = t.$$

For this equation to hold we must have that

$$\alpha = -1/2 \qquad \text{and} \qquad L(t) \sim 1/2 \qquad \text{or} \qquad L(t) \sim -\frac{1}{2},$$

so that

$$g''(t) \sim \frac{1}{2}t^{-1/2} \qquad \text{or} \qquad g''(t) \sim -\frac{1}{2}t^{-1/2}$$

and

$$g(t) \sim \frac{2}{3}t^{3/2} \qquad \text{or} \qquad g(t) \sim -\frac{2}{3}t^{3/2}.$$

The fact that we get two different possible asymptotics is encouraging, since we expect to get a two-dimensional space of solutions to our ODE. Before moving on, we can also check that other values of α do not yield anything sensible: If $\alpha = -1$ then g' is slowly varying and $g'' + (g')^2 = \pm o(t)$, while if $\alpha < -1$ then $g' = g'(\infty) \pm o(1)$ and $g'' + (g')^2 = \pm O(1) = \pm o(t)$. Thus, we have shown that **if g'' is regularly varying** then we must have that $g \sim \frac{2}{3}t^{3/2}$ or $g \sim -\frac{2}{3}t^{3/2}$ as $t \to +\infty$. This suggests we look for series expansions to solutions of $g'' + (g')^2 = t$ of the form

$$g = \frac{2}{3}t^{3/2} + ? + ? \ + \cdots$$

and

$$g = -\frac{2}{3}t^{3/2} + ? + ? \ + \cdots,$$

and we can guess what these missing functions should be by making more assumptions that relevant functions are regularly varying.

Let us start with the first solution, whose leading term is $\frac{2}{3}t^{3/2}$. To proceed, we will write our solution as

$$g(t) = \frac{2}{3}t^{3/2} + h(t)$$

for some function $h = g - \frac{2}{3}t^{3/2}$. The function h must satisfy

$$\left(\frac{2}{3}t^{3/2} + h(t)\right)'' + \left(\left(\frac{2}{3}t^{3/2} + h(t)\right)'\right)^2 = t,$$

so that

$$\frac{1}{2}t^{-1/2} + h'' + t + 2t^{1/2}h' + (h')^2 = t.$$

and hence that

$$h'' + 2t^{1/2}h' + (h')^2 = -\frac{1}{2}t^{-1/2}.$$

To proceed, we will **assume** that $h(t) = \pm o(t^{3/2})$ as $t \to +\infty$ and that $h'' = L(t)t^\alpha$ is regularly varying of some index α, with L a slowly varying function. Since $h = \pm o(t^{3/2})$ by assumption, we must have that $\alpha \leq -1/2$. As tends to be the case in these calculations, we need to do some case analysis according to the value of α. If $-1 < \alpha \leq -1/2$ then we have that

$$L(t)t^\alpha + \frac{2(1 \pm o(1))}{\alpha + 1}L(t)t^{\alpha+3/2} + \frac{(1 \pm o(1))}{(\alpha+1)^2}L(t)^2t^{2\alpha+2} = -\frac{1}{2}t^{-1/2}$$

To understand this equation we will have to split into further case analysis according to whether $\alpha < -1/2$ or $\alpha = -1/2$. In the first case, the term $L(t)t^{\alpha+3/2}$ dominates the left hand side and we have that

$$\frac{2(1 \pm o(1))}{\alpha + 1}L(t)t^{\alpha+3/2} = -\frac{1}{2}t^{-1/2},$$

which is incompatible with the assumption that $\alpha > -1$. If $\alpha = -1/2$ then $\alpha+3/2 = 2\alpha+2 = 1$ and we must have that

$$((4 \pm o(1))L(t) + (4 \pm o(1))L(t)^2)t = (L(t) - \frac{1}{2})t^{-1/2}.$$

Since the powers of t are different on the left is bigger than that on the right, this is only possible if the two terms $2(1 \pm o(1))L(t) + (1 \pm o(1))L(t)^2$ approximately cancel, which means that $L(t) \sim -L(t)^2$ and hence that $L(t) \sim -1$. But in this case $h'' \sim -t^{-1/2}$ and $h \sim -\frac{4}{3}t^{3/2}$, so that $|h|$ is not $o(t^{3/2})$. Indeed, we have just recovered the other "solution"

$$\frac{2}{3}t^{3/2} - \frac{4}{3}t^{3/2} = -\frac{2}{3}t^{3/2}.$$

We can also rule out the case that $\alpha = -1$, since in this case h' is either converging to a non-zero constant or slowly varying so that the term $2t^{1/2}h'$ is regularly varying with a larger index than any other term and the equation cannot hold.

It remains finally to consider the case $\alpha < -1$, in which case

$$h'(+\infty) - h'(t) = \int_t^\infty h''(s)\,\mathrm{d}s \sim \frac{L(t)t^{\alpha+1}}{|\alpha+1|},$$

where $h'(+\infty) = \lim_{t\to\infty} h'(t)$, so that

$$L(t)t^\alpha + 2h'(+\infty)t^{1/2} - \frac{2 \pm o(1)}{|\alpha+1|}L(t)t^{\alpha+3/2} + h'(+\infty)^2 \\ - \frac{2 \pm o(1)}{|\alpha+1|}h'(+\infty)L(t)t^{\alpha+1} + \frac{1 \pm o(1)}{|\alpha+1|^2}L(t)^2t^{2\alpha+2} = -\frac{1}{2}t^{-1/2}.$$

If $h'(+\infty) \neq 0$ then the term $2h'(+\infty)t^{1/2}$ has larger index than every other term, so that the equation cannot hold. Thus, we must have that $h'(+\infty) = 0$, in which case the equation simplifies to

$$L(t)t^\alpha - \frac{2 \pm o(1)}{|\alpha+1|}L(t)t^{\alpha+3/2} + \frac{1 \pm o(1)}{|\alpha+1|^2}L(t)^2t^{2\alpha+2} = -\frac{1}{2}t^{-1/2}.$$

Since $\alpha < -1$, the term involving $t^{\alpha+3/2}$ has larger index than anything else on the left hand side, so that

$$-\frac{2 \pm o(1)}{|\alpha+1|}L(t)t^{\alpha+3/2} = -\frac{1}{2}t^{-1/2}.$$

Thus, we must have that

$$\alpha = -2 \qquad \text{and } L(t) \sim \frac{1}{4} \qquad \text{as } t \to +\infty.$$

Thus, we have shown that **if** there is a solution to our ODE of the form $\frac{2}{3}t^{3/2} + h(t)$ such that $|h(t)| = o(t^{3/2})$ and h'' is regularly varying as $t \to +\infty$, then we must have that $h'' \sim \frac{1}{4}t^{-2}$ as $t \to +\infty$ and $h'(+\infty) = 0$, in which case

$$h'(t) \sim -\frac{1}{4}t^{-1} \qquad \text{as } t \to +\infty$$

and

$$h(t) \sim -\frac{1}{4}\log t \qquad \text{as } t \to +\infty.$$

This gives the next term in what will eventually be our infinite asymptotic series

$$g(t) = \frac{2}{3}t^{3/2} - \frac{1}{4}\log t + ? + ? \ + \ \cdots$$

One can perform a similar calculation for the other "solution", whose first term is $-\frac{2}{3}t^{3/2}$, and find that the next term must also be $-\frac{1}{4}\log t$, so that this other solution should be of the form

$$g(t) = -\frac{2}{3}t^{3/2} - \frac{1}{4}\log t + ? + ? \ + \ \cdots$$

What about the third term? Well, we can just do the same thing again: We assume that our solution is of the form

$$g(t) = \frac{2}{3}t^{3/2} - \frac{1}{4}\log t + h(t)$$

for some h satisfying $|h(t)| = o(\log t)$ as $t \to \infty$. Expanding out the ODE $g'' + (g')^2 = t$, we see that h must satisfy the ODE

$$\frac{1}{2}t^{-1/2} + \frac{1}{4}t^{-2} + h'' + t - \frac{1}{2}t^{-1/2} + \frac{1}{16}t^{-2} + 2(t^{1/2} + \log t)h' + (h')^2 = t$$

which we can simplify to

$$h'' + 2(t^{1/2} + \log t)h' + (h')^2 = -\frac{5}{16}t^{-2}.$$

To proceed, we can again assume that h'' is regularly varying of some index α, so that $h''(t) = L(t)t^{\alpha}$ for some slowly varying function $L(t)$. Since $h(t) = \pm o(t)$ we cannot have that $h'(t)$ converges to a non-zero constant, so that $h'(t)$ is regularly varying of index $\alpha + 1$. If $\alpha > -2$ then h must be regularly varying of index $\alpha + 2 > 0$, which is not possible since $h = \pm o(\log t)$, so that $\alpha \leq -2$. The term $2t^{1/2}h'$ has larger index than every other term on the left, so that

$$2t^{1/2}h' \sim -\frac{5}{16}t^{-2} \qquad \text{and} \qquad h(+\infty) - h(t) \sim -\frac{5}{48}t^{-3/2}.$$

Now, notice that adding a constant to g does not affect whether or not it solves $g'' + (g')^2 = t$, so that we should expect the freedom to choose a constant term, which corresponds to $h(+\infty)$.

(In the original linear Stoke's equation $f'' = tf$ this corresponds to multiplying our solution by e^A.) Thus, our asymptotic expansion should continue

$$g(t) = \frac{2}{3}t^{3/2} - \frac{1}{4}\log t + A + \frac{5}{48}t^{-3/2} + ? \ + \ \cdots$$

for a constant A that we are free to choose. Again, we can do something similar for the other "solution", this time obtaining that

$$g(t) = -\frac{2}{3}t^{3/2} - \frac{1}{4}\log t + A - \frac{5}{48}t^{-3/2} + ? \ + \ \cdots$$

where A is again a constant we are free to choose.

If we continue doing this for more and more terms, we will see that the next terms we get are constant multiples of t^{-3}, $t^{-9/2}$, t^{-6} etc. As such, it makes sense to try to find formal series solutions of the forms

$$g(t) = \frac{2}{3}t^{3/2} - \frac{1}{4}\log t + A + \sum_{n=2}^{\infty} a_n t^{(3-3n)/2}$$

and

$$g(t) = -\frac{2}{3}t^{3/2} - \frac{1}{4}\log t + A + \sum_{n=2}^{\infty} b_n t^{(3-3n)/2}.$$

Once we have identified the correct form for our formal solution, we can proceed in much the same way as when we dealt with formal power series. For the first form, we have the formal derivatives

$$\begin{aligned} g' &= t^{1/2} - \frac{1}{4t} + \sum_{n=2}^{\infty} \frac{3-3n}{2} a_n t^{(1-3n)/2} \\ &= \sum_{n=0}^{\infty} \left(\mathbb{1}(n=0) - \frac{1}{4}\mathbb{1}(n=1) + \frac{3-3n}{2} a_n \mathbb{1}(n \geq 2) \right) t^{(1-3n)/2} \\ &= t \sum_{n=0}^{\infty} \left(\mathbb{1}(n=0) - \frac{1}{4}\mathbb{1}(n=1) + \frac{3-3n}{2} a_n \mathbb{1}(n \geq 2) \right) \left(t^{-3/2} \right)^n \end{aligned}$$

and

$$\begin{aligned} g'' &= \sum_{n=0}^{\infty} \frac{1-3n}{2} \left(\mathbb{1}(n=0) - \frac{1}{4}\mathbb{1}(n=1) + \frac{3-3n}{2} a_n \mathbb{1}(n \geq 2) \right) t^{(-1-3n)/2} \\ &= \frac{1}{t} \sum_{n=0}^{\infty} \frac{1-3n}{2} \left(\mathbb{1}(n=0) - \frac{1}{4}\mathbb{1}(n=1) + \frac{3-3n}{2} a_n \mathbb{1}(n \geq 2) \right) \left(t^{-3/2} \right)^n. \end{aligned}$$

Thus, the formal equation $g'' + (g')^2 = t$ can be rewritten

$$\begin{aligned}
&\frac{1}{t}\sum_{n=0}^{\infty}\frac{1-3n}{2}\left(\mathbb{1}(n=0)-\frac{1}{4}\mathbb{1}(n=1)+\frac{3-3n}{2}a_n\mathbb{1}(n\geq 2)\right)\left(t^{-3/2}\right)^n \\
&\quad + t^2\sum_{n=0}^{\infty}\left[\sum_{k=0}^{n}\left(\mathbb{1}(k=0)-\frac{1}{4}\mathbb{1}(k=1)+\frac{3-3k}{2}a_k\mathbb{1}(k\geq 2)\right)\right. \\
&\qquad \left.\cdot\left(\mathbb{1}(n-k=0)-\frac{1}{4}\mathbb{1}(n-k=1)+\frac{3-3(n-k)}{2}a_{n-k}\mathbb{1}(n-k\geq 2)\right)\right]\left(t^{-3/2}\right)^n \\
&\qquad\qquad = t.
\end{aligned}$$

As a formal equality, this means that the coefficient of each power of t is equal on both sides. (Unlike before, however, we now have non-integer powers of t appearing in the series.) One can show that this does have a solution by finding a recursion for the coefficients a_n as we have done several times before. Rather than doing this, we will instead go back to our original equation, the Stokes equation $f'' = tf$.

Our (non-rigorous!!) calculations above suggest that we should have two solutions to the Stokes equation of the forms

$$f(t) = \exp\left[\frac{2}{3}t^{3/2} - \frac{1}{4}\log t + A + \sum_{n=2}^{\infty} a_n t^{(3-3n)/2}\right]$$

and

$$f(t) = \exp\left[-\frac{2}{3}t^{3/2} - \frac{1}{4}\log t + A + \sum_{n=2}^{\infty} b_n t^{(3-3n)/2}\right]$$

in some appropriate asymptotic sense. Since the exponential of a series in powers of $t^{-3/2}$ should itself be a series in powers of $t^{-3/2}$, it therefore makes sense to look for (formal) solutions to the Stokes equation of the form

$$f(t) = \exp\left[\frac{2}{3}t^{3/2} - \frac{1}{4}\log t\right]\sum_{n=0}^{\infty} a_n\left(t^{-3/2}\right)^n = \frac{e^{\frac{2}{3}t^{3/2}}}{t^{1/4}}\sum_{n=0}^{\infty} a_n\left(t^{-3/2}\right)^n$$

and

$$f(t) = \exp\left[-\frac{2}{3}t^{3/2} - \frac{1}{4}\log t\right]\sum_{n=0}^{\infty} b_n\left(t^{-3/2}\right)^n = \frac{e^{-\frac{2}{3}t^{3/2}}}{t^{1/4}}\sum_{n=0}^{\infty} b_n\left(t^{-3/2}\right)^n,$$

where the sequences a and b are **not** the same as in the expansion of the exponential. (I just want to avoid introducing more letters!) Let us focus on the first solution. To be a formal solution, all coefficients of $e^{\frac{2}{3}t^{3/2}}t^{\alpha}$ on both sides of the equation $f'' = tf$ should be equal to zero when we formally differentiate (i.e., write out the formal series in which we differentiate

term by term). The formal derivative of the series is

$$\begin{aligned}\sum_{n=0}^{\infty} a_n \frac{d}{dt}\left[\frac{e^{\frac{2}{3}t^{3/2}}}{t^{1/4}}\left(t^{-3/2}\right)^n\right] &= \sum_{n=0}^{\infty} a_n \frac{d}{dt}\left[e^{\frac{2}{3}t^{3/2}} t^{-3n/2-1/4}\right] \\ &= \sum_{n=0}^{\infty} a_n \left(e^{\frac{2}{3}t^{3/2}} t^{-3n/2+1/4} - \frac{6n+1}{4} e^{\frac{2}{3}t^{3/2}} t^{-3n/2-5/4}\right) \\ &= \frac{e^{\frac{2}{3}t^{3/2}}}{t^{1/4}} \sum_{n=0}^{\infty} a_n \left(t^{-3n/2+1/2} - \frac{6n+1}{4} t^{-3n/2-1}\right)\end{aligned}$$

and the formal second derivative is

$$\begin{aligned}\sum_{n=0}^{\infty} a_n \frac{d^2}{dt^2}\left[\frac{e^{\frac{2}{3}t^{3/2}}}{t^{1/4}}\left(t^{-3/2}\right)^n\right] &= \sum_{n=0}^{\infty} a_n \frac{d}{dt}\left(e^{\frac{2}{3}t^{3/2}} t^{-3n/2+1/4} - \frac{6n+1}{4} e^{\frac{2}{3}t^{3/2}} t^{-3n/2-5/4}\right) \\ &= \sum_{n=0}^{\infty} a_n \Bigg(e^{\frac{2}{3}t^{3/2}} t^{-3n/2+3/4} - \frac{6n-1}{4} e^{\frac{2}{3}t^{3/2}} t^{-3n/2-3/4} \\ &\qquad - \frac{6n+1}{4} e^{\frac{2}{3}t^{3/2}} t^{-3n/2-3/4} + \frac{6n+1}{4}\frac{6n+5}{4} e^{\frac{2}{3}t^{3/2}} t^{-3n/2-9/4}\Bigg) \\ &= \frac{e^{\frac{2}{3}t^{2/3}}}{t^{1/4}} \sum_{n=0}^{\infty} a_n \Bigg(t^{-3n/2+1} - \frac{6n-1}{4} t^{-3n/2-1/2} \\ &\qquad - \frac{6n+1}{4} t^{-3n/2-1/2} + \frac{6n+1}{4}\frac{6n+5}{4} t^{-3n/2-2}\Bigg).\end{aligned}$$

Thus, the equation $f'' = tf$ becomes

$$\frac{e^{\frac{2}{3}t^{2/3}}}{t^{1/4}} \sum_{n=0}^{\infty} a_n \left(t^{-3n/2+1} - 3nt^{-3n/2-1/2} + \frac{6n+1}{4}\frac{6n+5}{4} t^{-3n/2-2}\right) = \frac{e^{\frac{2}{3}t^{3/2}}}{t^{1/4}} \sum_{n=0}^{\infty} a_n t^{-3n/2+1}$$

which is equivalent to

$$\sum_{n=0}^{\infty} 3n a_n t^{-3n/2} = \sum_{n=0}^{\infty} a_n \frac{6n+1}{4}\frac{6n+5}{4} t^{-3(n+1)/2} = \sum_{n=0}^{\infty} a_{n-1} \frac{6n-5}{4}\frac{6n-1}{4} t^{-3n/2} \mathbb{1}(n \geq 1),$$

so that

$$a_n = \frac{(6n-1)(6n-5)}{48n} a_{n-1} = \cdots = a_0 \prod_{i=1}^{n} \frac{(6i-1)(6i-5)}{48i} \tag{8.9}$$

for every $n \geq 1$. Since the ratio $a_n/a_{n-1} \sim 36n/48 \to \infty$, the formal series will have infinite radius of convergence.

The standard way to express the coefficients is in terms of the **Gamma function** $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, \mathrm{d}t$, which has $\Gamma(1) = 1$ and (using integration by parts) has the property that

$\Gamma(x+1) = x\Gamma(x)$ for every $x > 0$. This means that $n! = \Gamma(n+1)$ for each integer $n \geq 0$, so that the Gamma function may be thought of as a natural continuous version of the factorial function.

Exercise 71. Prove that if $a > -1$ then

$$\prod_{i=1}^{n}(i+a) = \frac{\Gamma(m+a+1)}{\Gamma(a+1)}$$

for every $n \geq 0$. Deduce that the sequence (a_n) appearing in (8.9) can be written

$$a_n = \left(\frac{3}{4}\right)^n \frac{\Gamma(n+5/6)\Gamma(n+1/6)}{\Gamma(5/6)\Gamma(1/6)\Gamma(n+1)} a_0$$

for every $n \geq 0$.

One can do a similar calculation for the other solution, giving that

$$b_n = (-1)^n \left(\frac{3}{4}\right)^n \frac{\Gamma(n+5/6)\Gamma(n+1/6)}{\Gamma(5/6)\Gamma(1/6)\Gamma(n+1)} b_0,$$

In particular, one sees again that the series $\sum_{n=0}^{\infty} b_n t^{-3n/2}$ does not converge for any $t \in \mathbb{R}$, so that these series solutions can be interpreted only as asymptotic expansions. All this work has led us to divergent formal series solutions of the form

$$\frac{e^{\frac{2}{3}t^{3/2}}}{t^{1/4}} \sum_{n=0}^{\infty} \left(\frac{3}{4}\right)^n \frac{\Gamma(n+5/6)\Gamma(n+1/6)}{\Gamma(5/6)\Gamma(1/6)\Gamma(n+1)} t^{-3n/2}$$

and

$$\frac{e^{-\frac{2}{3}t^{3/2}}}{t^{1/4}} \sum_{n=0}^{\infty} (-1)^n \left(\frac{3}{4}\right)^n \frac{\Gamma(n+5/6)\Gamma(n+1/6)}{\Gamma(5/6)\Gamma(1/6)\Gamma(n+1)} t^{-3n/2}.$$

Of course we would really like to say that we really have solutions to the Stokes equation satisfying

$$f(t) = a_0 \sum_{n=0}^{\infty} \left(\frac{3}{4}\right)^n \frac{\Gamma(n+5/6)\Gamma(n+1/6)}{\Gamma(5/6)\Gamma(1/6)\Gamma(n+1)} \frac{e^{\frac{2}{3}t^{3/2}}}{t^{1/4}} t^{-3n/2} \quad \text{mod t.s.t.s} \qquad \text{as } t \to +\infty$$

and

$$f(t) = b_0 \sum_{n=0}^{\infty} (-1)^n \left(\frac{3}{4}\right)^n \frac{\Gamma(n+5/6)\Gamma(n+1/6)}{\Gamma(5/6)\Gamma(1/6)\Gamma(n+1)} \frac{e^{-\frac{2}{3}t^{3/2}}}{t^{1/4}} t^{-3n/2} \quad \text{mod t.s.t.s} \qquad \text{as } t \to +\infty.$$

While there are good general ways to prove this last step, we only have time to discuss them very briefly in this course. Let us just say that this is correct, and that in fact there are two linearly independent solutions to the Stokes equation satisfying these asymptotics that can

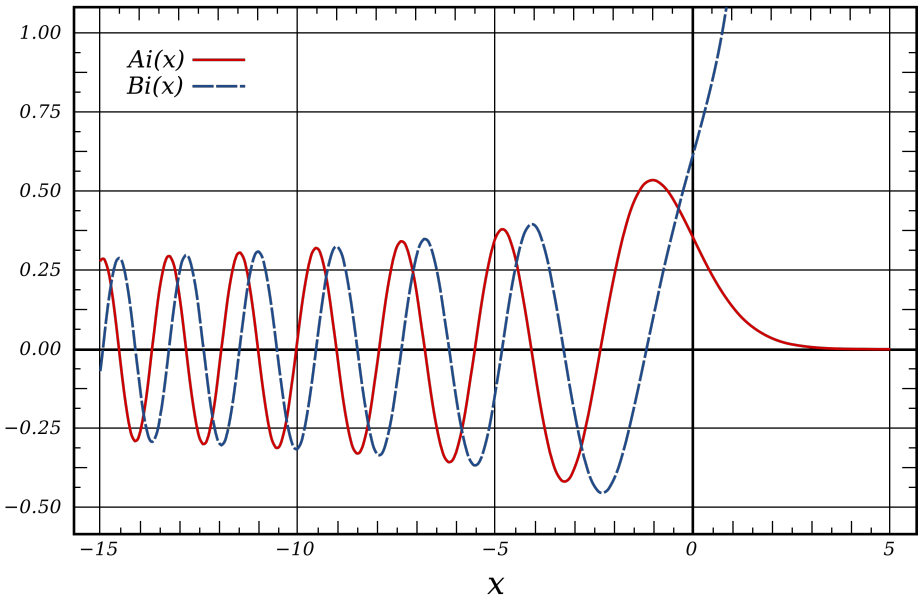


Figure 8: Graphs of the Airy functions $\mathrm{Ai}(x)$ and $\mathrm{Bi}(x)$. (Courtesy of Wikipedia.)

be written

$$\mathrm{Bi}(t) = \frac{1}{\pi}\int_0^\infty \left[\exp\left(-\frac{x^3}{3} + tx\right) + \sin\left(\frac{x^3}{3} + tx\right)\right] \mathrm{d}x$$

and

$$\mathrm{Ai}(t) = \frac{1}{\pi}\int_0^\infty \cos\left(\frac{x^3}{3} + tx\right) \mathrm{d}x.$$

The functions $\mathrm{Ai}(t)$ and $\mathrm{Bi}(t)$ are known as the **Airy functions of the first and second kind**. (I apologize for using the sequence $(a_n)_{n\geq 0}$ for the coefficients of $\mathrm{Bi}(t)$ and $(b_n)_{n\geq 0}$ for the coefficients of $\mathrm{Ai}(t)$!)

Remark 8.9 (Stokes phenomenon). The functions $\mathrm{Ai}(t)$ and $\mathrm{Bi}(t)$ both oscillate between positive and negative values as $t \to -\infty$, and the asymptotic expansions we wrote down for them are *not* valid as asymptotic expansions in the $t \to -\infty$ limit. Indeed, the correct first order asymptotics of $\mathrm{Ai}(t)$ as $t \to -\infty$ turn out to be

$$\mathrm{Ai}(-t) = \frac{C_1 \pm o(1)}{t^{1/4}} \sin\left(\frac{2}{3}t^{3/2} + \frac{\pi}{4}\right) - \frac{C_2 \pm o(1)}{t^{1/4}} \cos\left(\frac{2}{3}t^{3/2} + \frac{\pi}{4}\right)$$

for appropriate positive constants C_1 and C_2, and it is possible to give a full asymptotic expansion in which these $\pm o(1)$ terms are expanded in powers of $t^{-3/2}$. Now consider the function

$$f(t) = \begin{cases} t^2\,\mathrm{Ai}(\frac{1}{t}) & x \neq 0 \\ 0 & x = 0. \end{cases}$$

which is twice differentiable with first and second derivatives equal to 0 at 0. For $t \neq 0$ we can compute that

$$\frac{d}{dt} t^2 \operatorname{Ai}(t^{-1}) = 2t \operatorname{Ai}(t^{-1}) - t \operatorname{Ai}'(t^{-1})$$

and that

$$\frac{d^2}{dt^2} t^2 \operatorname{Ai}(t^{-1}) = 2 \operatorname{Ai}(t^{-1}) + 2 \operatorname{Ai}'(t^{-1}) + \operatorname{Ai}''(t^{-1}) = (t+2) \operatorname{Ai}(t^{-1}) + 2 \operatorname{Ai}'(t^{-1})$$

where in the last step we used that $\operatorname{Ai}(t)$ solves Stokes equation $\operatorname{Ai}''(t) = t \operatorname{Ai}(t)$. This means that

$$\begin{aligned} t^3 f''(t) = (t^2+2t)f + 2t^2 \operatorname{Ai}'(t^{-1}) = (t^2+2t)f + 2t(2t \operatorname{A_i}(t^{-1}) - f') \\ = (t^2+2t)f + 4f - 2tf' \end{aligned}$$

so that f solves the ODE

$$t^3 f'' = (t^2 + 2t + 4)f - 2tf',$$

which is a polynomial ODE that is not of the standard form $f^{(m)} = P(t, f, \ldots, f^{(m-1)})$. Why is this interesting? Well, the fact that we needed different asymptotic expansions to understand the behaviour of $\operatorname{Ai}(t)$ near $+\infty$ and near $-\infty$ means that we need different asymptotic expansions to understand the solution f to this ODE as t approaches 0 from above and from below. This is the **Stokes phenomenon**: the solutions to an ODE may have different asymptotic behaviours as $t \to t_0$ depending on which direction one approaches the point t_0 from! Note that this cannot happen when the solution is analytic at t_0.

Exercise 72. Let f be a solution to the polynomial ODE $f'' - 2f' = f^2$.

1. Prove that if f'' is regularly varying as $t \to +\infty$ then $f \sim 2/t$ as $t \to +\infty$.
2. Prove that the ODE admits a formal series solution of the form

$$f = \sum_{n=1}^{\infty} \frac{P_n(\log t)}{t^n}$$

where $(P_n)_{n \geq 0}$ is a sequence of degree $n-1$ polynomials with leading coefficient 2 for each $n \geq 0$.

8.5 Tauberian and Abelian theory

In this section we will briefly sketch some answers to the following questions:

1. How can we extract asymptotic information about its function from its Laplace transform, even when we cannot invert it explicitly?

2. How can we extract asymptotic information about a sequence from its (ordinary or exponential) generating function, even when we cannot get an explicit formula for the sequence?

(Note in both cases that even if we do have an explicit formula, it might still be hard to estimate the asymptotic growth of this formula, in which case the techniques of this section may still be helpful.)

A theorem which lets us go from function / sequence asymptotics to Laplace transform / generating function asymptotics is called an **Abelian Theorem**; a theorem which lets us go from Laplace transform / generating function asymptotics to function / sequence asymptotics is called a **Tauberian Theorem**. Note that Tauberian/Abelian theorems always concern first-order asymptotics only, and typically do not yield explicit error bounds. The various Abelian and Tauberian theorems available differ in the hypotheses they place on the functions in order to deduce their conclusions, which in this context are often called **side conditions**. Tauberian theorems are typically more interesting/useful than Abelian theorems, and will be our focus in this section.

We begin by stating the Hardy-Littlewood-Karamata Tauberian Theorem, which allows us to extract asymptotic estimates for a **non-negative** sequence from its generating function. (Here non-negativity is the "side condition.") Karamata introduced the theory of regularly varying functions, which allowed him to generalize and simplify the earlier Tauberian theorem of Hardy and Littlewood.

Theorem 8.10 (Hardy-Littlewood-Karamata Tauberian Theorem)**.** *Let $f(t) = \sum_{n=0}^{\infty} a_n t^n$ be a power series with non-negative coefficients and radius of convergence $t_c > 0$. If*

$$f(t) \sim \left(\frac{t_c}{t_c - t}\right)^{\alpha} L\left(\frac{t_c}{t_c - t}\right)$$

as $t \uparrow t_c$ for some $\alpha \geq 0$ and some continuous slowly varying function L then

$$\sum_{n=0}^{N} a_n t_c^n \sim \frac{L(N) N^{\alpha}}{\Gamma(\alpha + 1)}$$

as $N \to \infty$.

Often there are some mild additional conditions that allow us to convert this into a pointwise estimate

$$a_n \sim \frac{L(n) n^{\alpha - 1}}{\Gamma(\alpha)} t_c^{-n}$$

as $n \to \infty$.

Exercise 73. Let $f(t) = \sum_{n=0}^{\infty} a_n t^n$ be a power series with non-negative coefficients and radius of convergence $t_c > 0$. Prove that if

$$f(t) \sim \left(\frac{t_c}{t_c - t}\right)^{\alpha} L\left(\frac{t_c}{t_c - t}\right)$$

as $t \uparrow t_c$ for some $\alpha \geq 0$ and some continuous slowly varying function L *and* $a_{n+1}t_c^{n+1} \geq a_n t_c^n$ *for every* $n \geq 0$ then

$$a_n \sim \frac{L(n)n^{\alpha-1}}{\Gamma(\alpha)} t_c^{-n}$$

as $N \to \infty$.

In order to apply this theorem, we will often need to also use the following important theorem about power series with non-negative coefficients.

Theorem 8.11 (Vivanti-Pringsheim Theorem)**.** *Let* $f(t) = \sum_{n=0} a_n t^n$ *be a power series with non-negative coefficients and radius of convergence* $r > 0$*. Then there is no real-analytic function defined on an interval* $(r-\varepsilon, r+\varepsilon)$ *that coincides with* f *on the interval* $(r-\varepsilon, r)$*.*

This theorem is often summarized by the slogan "functions defined by power series with non-negative coefficients have singularities on the positive real axis at their radius of convergence". Be careful to note that the relevant notion of singularity can be subtle, however, and does *not* imply that $|f(t)| \to \infty$ as $t \uparrow r$.

Example 8.12. Let $(A_n)_{n\geq 0}$ be the sequence defined in example 7.6, so that

$$\sum_{n=0}^{\infty} \frac{A_n}{n!} t^n = \tan\left(\frac{t}{2} + \frac{\pi}{4}\right)$$

for t within the radius of convergence of the left hand side. Since $A_n/n!$ is non-negative for every $n \geq 0$ and $\tan(t/2 + \pi/4)$ is analytic on $(-3\pi/2, \pi/2)$, the radius of convergence of the series on the left must be $\pi/2$ and the equality between the two functions must hold for all $-\pi/2 < t < \pi/2$ by Vivanti-Pringsheim and rigidity of real analytic functions. Since we also have that

$$\tan\left(\frac{t}{2} + \frac{\pi}{4}\right) = \frac{\sin\left(\frac{t}{2} + \frac{\pi}{4}\right)}{\cos\left(\frac{t}{2} + \frac{\pi}{4}\right)} \sim \frac{1}{\pi/4 - t/2} = \frac{4}{\pi} \frac{\pi/2}{\pi/2 - t}.$$

Applying the Hardy-Littlewood-Karamata Tauberian Theorem, it follows that

$$\sum_{n=0}^{N} \frac{A_n}{n!} \left(\frac{2}{\pi}\right)^n \sim \frac{4}{\pi} N$$

as $N \to \infty$.

Exercise 74. Prove that this sequence satisfies

$$A_n \sim \frac{4}{\pi} \left(\frac{\pi}{2}\right)^n n!$$

as $n \to \infty$.

We now turn to the analogous theorem for the Laplace transform.

Theorem 8.13 (Karamata)**.** *Let $f : [0,\infty) \to [0,\infty)$ be a continuous function whose Laplace transform $\mathcal{L}\{f\}$ has domain (s^*,∞) for some $s^* \in \mathbb{R}$, and let $\phi : [0,\infty) \to (0,\infty)$ be regularly varying of index $\alpha \geq 0$. Then the following are equivalent.*

1. $\int_0^T e^{-s^*t} f(t)\,\mathrm{d}t \sim \phi(T)$ *as* $T \to +\infty$.
2. $\mathcal{L}\{f\}(s) \sim \Gamma(\alpha+1)\phi\left(\frac{1}{s-s^*}\right)$ *as* $s \downarrow s^*$.

Again, if $\phi(t) = t^\alpha L(t)$ for some slowly-varying function L, one can deduce under mild additional assumptions that one also has the pointwise asymptotic estimate

$$f(t) \sim \alpha e^{s^*t} L(t) t^{\alpha-1}.$$

(For example this holds if we make the additional assumption that f is regularly varying.)

Example 8.14. Earlier we studied the ODE $f'' - tf' - f = 0$, and showed that if a solution to this equation has Laplace transform whose domain includes $(0,\infty)$ then this Laplace transform must satisfy

$$\mathcal{L}\{f\}(s) = e^{-\frac{1}{2}(s^2-1^2)}\left(\mathcal{L}\{f\}(1) + \int_1^s e^{\frac{1}{2}(u^2-1^2)}(f(0+) + \frac{1}{u}f'(0+))\,\mathrm{d}u\right)$$

for every $s > 0$. If $f'(0+)$ then we have that

$$\mathcal{L}\{f\}(s) \sim -\int_s^1 \frac{f'(0+)}{u}\,\mathrm{d}u = -f'(0+)\log\frac{1}{s}$$

as $s \downarrow 0$, and it follows from Karamata's theorem (applied to $-f$) that if this holds and f is non-positive then

$$\int_0^T f(t)\,\mathrm{d}t \sim -f'(0+)\log T$$

as $T \to +\infty$. It follows moreover that if f is regularly varying then

$$f(t) \sim -\frac{f'(0+)}{t}$$

as $t \to +\infty$.

Higher-order Tauberian analysis

Tauberian theory allows us to draw conclusions with only very light assumptions on our functions, but tends to only yield first-order asymptotics and does not provide error estimates. In the case that our Laplace transform/generating function is 'nice' and we can do more computations with it, there are often much more powerful tools available that can give us the whole asymptotic expansion (which may be a convergent expansion). A very detailed account is given in the textbook *Analytic Combinatorics* by Flajolet and Sedgewick. One of the most sophisticated versions of this analysis, which is designed to handle the case that the generating function has singularities at every (complex) point in the boundary of its disc of convergence, is the *Hardy-Ramanujan-Littlewood Circle Method*. The circle method was famously applied by Hardy and Ramanujan to analyze the number of ways to write a number n as an *unordered* sum of positive integers, known as the *partition number* $p(n)$, which (setting $p(0) = 1$) has ordinary generating function expressible as an infinite product

$$\sum_{n=0}^{\infty} p(n)t^n = \frac{1}{\prod_{i=1}^{\infty}(1-t^i)} \text{ for } |t| < 1.$$

Hardy and Ramanujan used their very sophisticated version of Tauberian theory to prove that

$$p(n) \sim \frac{\sqrt{3}}{12n} e^{\pi\sqrt{\frac{2n}{3}}} \qquad \text{as } n \to \infty,$$

and also to get very precise control of the errors in this approximation. Their paper (published 1918) is remarkably modern – I highly recommend that you read the original if you are interested!

Dirichlet series and the prime number theorem

Another very important application of Tauberian theory in math is the *prime number theorem*, which states that if $\pi(n)$ is the number of primes smaller than n then

$$\pi(n) \sim \frac{n}{\log n}$$

as $n \to \infty$. Rather than ordinary or exponential generating functions, the proof of the prime number theorem relies on the analysis of *Dirichlet series* $\sum_{n=1}^{\infty} \frac{a_n}{n^s}$, a kind of generating function that appears very naturally in number-theoretic contexts. The relevance of these series to number theory comes from their multiplication law:

$$\left(\sum_{n=1}^{\infty} \frac{a_n}{n^s}\right)\left(\sum_{n=1}^{\infty} \frac{b_n}{n^s}\right) = \sum_{n=0}^{\infty}\sum_{n=1}^{\infty} \frac{\sum_{k\ell=n} a_k b_\ell}{n^s}$$

The operation $a, b \mapsto a * b$ with $(a * b)_n := \sum_{k\ell=n} a_k b_\ell$ is known as **Dirichlet convolution**. The importance of the Riemann zeta function $\zeta(s) := \sum_{n=1}^{\infty} n^{-s}$ owes in part to its being the Dirichlet series of the all-one sequence, so that replacing a sequence $(a_n)_{n\geq 1}$ by the sum $(\sum_{m \text{ divides } n} a_m)_{n\geq 1}$ corresponds to multiplying the Dirichlet series of a by the Riemann zeta function.

9 A brief introduction to dynamical systems

For the last part of the course, we will consider a completely different approach to ODEs, in which we consider the *qualitative* properties of solutions to autonomous differential equations. Here, we recall that a differential equation is autonomous if it does not directly involve the time variable t. We will restrict attention to equations of the form $f^{(m)} = F(f, f', \dots, f^{(m-1)})$. Since every autonomous differential equation of the form is equivalent to a *first order* autonomous equation $\mathbf{f}' = F(\mathbf{f})$ in higher dimensions, it suffices to consider first order autonomous equations. We will restrict further to the case that F is continuously differentiable and defined on all of $\mathbb{R}^d$, so that by global Picard-Lindelöf there is a unique solution maximal solution with $\mathbf{f}(0) = x$ for each $x \in \mathbb{R}^d$.

As we noted when we previously discussed autonomous equations, if $(I, \mathbf{f}(t))$ is a solution to the autonomous equation $\mathbf{f}' = F(\mathbf{f})$ then so is the shifted function $(I + t_0, \mathbf{f}(t - t_0))$ for each $t_0 \in \mathbb{R}$. As such, we can think of the ODE $\mathbf{f}' = F(\mathbf{f})$ as partitioning $\mathbb{R}^d$ into a collection of curves, where two points x and y lie on the same curve if there is a maximal solution to the ODE passing through both x and y: If such a curve exists, then it is uniquely specified up to time-shifts. These curves are called the *trajectories* of the autonomous ODE, and the study of the partition of $\mathbb{R}^d$ into the trajectories of the system is known as *dynamics*.

It is often helpful to think of the function F in the ODE $\mathbf{f}' = F(\mathbf{f})$ as a *vector field*, in which each element of x is assigned to a vector $F(x)$. The vector $F(x)$ describes the way a trajectory started at x will move by

$$f(t) = x + F(x)t \pm o(|t|)$$

as $t \to 0$, where f is the solution to the ODE with $f(0) = x$. As such, to understand the *shape* of the trajectories of the ODE associated to F, we only need to know 1) the points where F is equal to 0, which will yield constant solutions $f(t) \equiv x$ – these are called the **fixed points** or **equilibrium points** of the ODE, and 2) the *direction* of $F(x)$ when $F(x) \neq 0$, i.e., the unit vector $F(x)/\|F(x)\|$. Indeed, if F and $\tilde{F}$ are two vector fields related by $\tilde{F}(x) = \lambda(x)F(x)$ for some function $\lambda : \mathbb{R}^d \to (0, \infty)$ then $\tilde{F}$ and F will decompose $\mathbb{R}^d$ into the same set of trajectories (but with possibly different time parameterizations of associated solutions). As such, we often represent vector fields in $\mathbb{R}^2$ graphically by drawing arrows of unit length in the direction of the vector field at each non-equilibrium point. The trajectories (a.k.a. flow lines) of the vector field just follow these arrows!

We already know how to understand one-dimensional linear autonomous ODEs since they are always separable. Indeed, away from equilibrium points we have an implicit solution to the ODE $f' = F(f)$ given by

$$\int \frac{\mathrm{d}f}{F(f)} = \int 1 \, \mathrm{d}t,$$

which we can always invert to get a solution away from equilibrium points since $\int \frac{\mathrm{d}f}{F(f)}$ is

monotone in f whenever F is continuous and does not take the value 0.

Stable and unstable equilibria. Let $F : \mathbb{R} \to \mathbb{R}$ be continuously differentiable, and suppose that x_0 is an equilibrium point of $f' = F(f)$, so that $F(x_0) = 0$.

1. We say that x_0 is an **asymptotically stable equilibrium** (a.k.a. attractive fixed point) if there exists $\varepsilon > 0$ such that if $|x - x_0| \leq \varepsilon$ then the trajectory starting at x converges to x_0 as $t \to +\infty$.

2. We say that x_0 is a **stable equilibrium** if for each $\varepsilon > 0$ there exists $\delta > 0$ such that if $|x - x_0| \leq \delta$ then the trajectory starting at x stays within the ball of radius ε around x_0 for all $t \geq 0$.

3. We say that x_0 is an **unstable equilibrium** (a.k.a. repulsive fixed point) if there exists a positive constant c such that any solution f started at a point $x \neq x_0$ satisfies $|f(t) - x_0| \geq c$ for some $t \geq 0$.

Intuitively, if we start near an asymptotically stable equilibrium our solution will be 'sucked in' and get closer to the equilibrium point over time, while at an unstable equilibrium point, no matter how close we start to the point, we will eventually get far away from the point. A stable equilibrium may not be an asymptotically stable equilibrium if the error $f(t) - x_0$ tends to stay of the same order rather than shrink over time.

Proposition 9.1. *Let x_0 be an equilibrium point of a one-dimensional autonomous ODE $f' = F(f)$ with F continuously differentiable. If $F'(x_0) < 0$ then x_0 is asymptotically stable, while if $F'(x_0) > 0$ then x_0 is unstable.*

Proof. The more basic criterion is that

1. If there exists $\varepsilon > 0$ such that $F(x)$ is positive for $x < x_0$ with $|x - x_0| \leq \varepsilon$ and $F(x)$ is negative for $x > x_0$ with $|x - x_0| \leq \varepsilon$ then x_0 is an asymptotically stable fixed point.

2. If there exists $\varepsilon > 0$ such that $F(x)$ is negative for $x < x_0$ with $|x - x_0| \leq \varepsilon$ and $F(x)$ is positive for $x > x_0$ with $|x - x_0| \leq \varepsilon$ then x_0 is an unstable fixed point.

The derivative conditions in the proposition are just sufficient conditions for these to hold.

We will prove the stability criterion, leaving the instability criterion as an exercise. We will prove the stronger, 'one-sided' version of this claim that if $F(x) < 0$ for $x \in (x_0, x_0 + c)$ for some $c > 0$, then any solution started at $x \in (x_0, x_0 + c)$ must converge to x_0 as $t \to +\infty$. Let f be the solution to $f' = F(f)$ with $f(0) = x$ and let $T = \inf\{t : f(t) \notin (x_0, x_0 + c)\}$, where we set $\inf \emptyset = +\infty$. We wish to show that $T = \infty$ and that $f(t) \to x_0$ as $t \to +\infty$. Since $f' = F(f)$, the derivative of f is negative on $[0, T)$ and we must have by the mean-value theorem that f is decreasing on $[0, T)$. On the other hand, we have by global Picard-Lindelöf that the constant function $g \equiv x_0$ is the *only* solution to our ODE passing through x_0, so there cannot be a point in the domain of f with $f(t) = x_0$. As such, we must have that $T = \infty$, that $f(t) \in (x_0, x]$ for every $t \geq 0$, and that $f(t)$ is a decreasing function of t.

Exercise 75. Prove that the domain of the solution f includes $[0, \infty)$.

Now, since f is decreasing and bounded below, it must converge to some limit $f(+\infty)$ as $t \to +\infty$. Since $f(t) = f(0) + \int_0^t f'(s)\,\mathrm{d}s$, this is inconsistent with f' converging to a non-zero constant, and since F is continuous we must have that $F(f(+\infty)) = 0$. Since $f(t) \in (x_0, x]$ for every $t \geq 0$ and F is non-zero at every point of $(x_0, x]$, this implies that $f(+\infty) = x_0$ as claimed. □

Exercise 76. Prove the instability part of the proposition.

At a more intuitive level, the idea is that when $F'(x_0) \neq 0$ and $f(0)$ is close to x_0, the function f satisfies

$$f'(t) = F(f(t)) = F'(x_0)f(t) \pm o(|f(t)|)$$

so that f should we well-approximated at small times by the solution to the *linear*, constant coefficient ODE

$$g' = F'(x_0)g$$

with $g(0) = f(0)$. Using this approximation to understand the behaviour of an ODE near an equilibrium point is called *linearization around the fixed point.* Note that while this approximation is still valid in some senses when $F'(x_0) = 0$, it is not sufficient to determine whether the point is stable or unstable. (Note however that near a stable equilibrium point with $F'(x_0) = 0$, the solution will be sucked into x_0 much more slowly than it will be at a stable equilibrium point with $F'(x_0) < 0$. Similarly, solutions started near an unstable equilibrium point with $F'(x_0) = 0$ will escape from the point much more slowly than at an equilibrium point with $F'(x_0) > 0$.)

If $F'(x_0) = 0$ then we cannot determine whether the point is stable or unstable by looking at the derivative alone, and must investigate some higher-order information.

Example 9.2.

1. $f' = F(f) = f^3$ has an unstable equilibrium at zero with $F'(0) = 3(0)^2 = 0$.
2. $f' = F(f) = -f^3$ has an asymptotically stable equilibrium at zero with $F'(0) = -3(0)^2 = 0$.
3. $f' = F(f) = f^2$ has an equilibrium at zero with $F'(0) = 2(0) = 0$ that is neither stable nor unstable: Solutions starting at small positive numbers escape to infinity, while solutions started at small negative numbers are sucked back in towards zero.

Let us note however that unless F has an infinite set of zeroes accumulating to the equilibrium points x_0, there is always a well-defined notion of x_0 being stable or unstable in each direction.

Similar reasoning to that used at the end of the previous proof leads to the following proposition.

Proposition 9.3. *Let (I, f) be a maximal solution to a one-dimensional autonomous ODE $f' = F(f)$ with F continuously differentiable. As $t \to \sup I$, $f(t)$ either converges to $+\infty$, $-\infty$, or an equilibrium point. The same holds as $t \to \inf I$. Moreover, the solution f cannot converge to an unstable equilibrium point as $t \to \sup I$ unless f starts at that point.*

Example 9.4 (The logistic equation)**.** Suppose that the number of cells in a bacterial colony is governed by the ODE

$$P'(t) = P(M - P)$$

for some constant M. If the ODE were just $P' = MP$ this would model pure exponential population growth. The term $(M - P)$ is supposed to model the effect of a large population using up resources so that growth slows and eventually becomes negative as the population increases. This ODE has two equilibrium points, 0 and M. Taking the P-derivative

$$\frac{d}{dP}P(M - P) = M - 2P,$$

we see that 0 is an unstable equilibrium point and that M is a stable equilibrium point. It follows that if P is a solution with $P(0) > 0$ then $P(t)$ converges to M as $t \to +\infty$, if $P(0) = 0$ then $P(t) = 0$ for every $t \geq 0$, while if $P(0) < 0$ then P converges to $-\infty$ as $t \to +\infty$.

Linearlization in higher dimensions. In higher dimensions one may also study the behaviour of an ODE $\mathbf{f} = F(\mathbf{f})$ near an equilibrium point x_0 by studying the linearized ODE $(\mathbf{f} - x_0)' = DF(x_0)(\mathbf{f} - x_0)$, but the situation is significantly more complicated than in one dimension since the linear map $DF(x_0)$ can have various different relevant properties besides being positive or negative. The simplest case to understand is when all eigenvalues have non-zero real part of constant sign.

Theorem 9.5. *If all the eigenvalues of $DF(x_0)$ have positive real part then x_0 is an unstable fixed point. If all the eigenvalues of $DF(x_0)$ have negative real part then x_0 is an asymptotically stable fixed point.*

Let us suppose for simplicity that $DF(x_0)$ is diagonalizable with real eigenvalues $\lambda_1, \ldots, \lambda_d$ and corresponding eigenvectors $e_1, \ldots, e_d$. Writing $f(t) = f_1(t)e_1 + \cdots + f_d(t)e_d$ and $x_0 = x_{0,1}e_1 + \cdots + x_{0,d}e_d$ we can compute that

$$\begin{aligned}\frac{d}{dt}\sum_{i=1}^{d}(f_i(t) - x_{0,i})^2 = \sum_{i=1}^{d}\frac{d}{dt}(f_i(t) - x_{0,i})^2 = \sum_{i=1}^{d}2f_i'(t)(f_i(t) - x_{0,i}) \\ = 2\sum_{i=1}^{d}F(f(t))_i(f_i(t) - x_{0,i}).\end{aligned}$$

Using the definition of the derivative, we can get with a little work that

$$\begin{aligned} 2\sum_{i=1}^{d} F(x)_i(x_i - x_{0,i}) &= 2\sum_{i=1}^{d} \left[DF(x_0)(x - x_0)\right]_i (x_i - x_{0,i}) \pm o(\|x - x_0\|_2^2) \\ &= 2\sum_{i=1}^{d} \lambda_i (x_i - x_{0,i})^2 \pm o(\|x - x_0\|_2^2) \end{aligned}$$

as $x \to x_0$, where we expand each vector x in the basis of eigenvectors $x = x_1 e_1 + \cdots x_d e_d$. If we assume that all the eigenvalues are positive, it follows that there exists $\varepsilon > 0$ such that if $f(t) \neq x_0$ and $\|f(t) - x_0\|_2 \leq \varepsilon$ then

$$\frac{d}{dt} \sum_{i=1}^{d} (f_i(t) - x_{0,i})^2 > 0.$$

Similarly, if all the eigenvalues are negative there exists $\varepsilon > 0$ such that if $f(t) \neq x_0$ and $\|f(t) - x_0\|_2 \leq \varepsilon$ then

$$\frac{d}{dt} \sum_{i=1}^{d} (f_i(t) - x_{0,i})^2 < 0.$$

This is can be seen to imply the claims about stability/instability following a similar argument to the one-dimensional case.

The technique of studying the solutions to multidimensional ODEs by finding functions $\phi : \mathbb{R}^d \to \mathbb{R}$ for which we can estimate the time derivative $\phi(f(t))'$ is very important. Given an ODE $f' = F(f)$ and an equilibrium point x_0, a continuously differentiable function $\phi : \mathbb{R}^d \to [0, \infty)$ is said to be a **Lyapunov function** if $\phi(x_0) = 0$ and there exists $\varepsilon > 0$ such that $\phi(x) > 0$ for every $x \neq x_0$ with $\|x - x_0\| \leq \varepsilon$, and $\frac{d}{dt}\phi(f(t)) \leq 0$ for any solution f starting with $\|f(0) - x_0\| \leq \varepsilon$. Any equilibrium point admitting a Lyapunov function must be stable, and if the function satisfies $\frac{d}{dt}\phi(f(t)) < 0$ for every solution starting at a point close to but not equal to x_0 then x_0 is asymptotically stable.

An equilibrium point x_0 is said to be **hyperbolic** if all of the eigenvalues of $DF(x_0)$ have non-zero real part. Around hyperbolic fixed points the dynamics of the system are always well-approximated by those of the linearization: When F is smooth a very strong statement to this effect is given by the *Hartman-Grobman theorem*, which states that for each hyperbolic fixed point there exists $\varepsilon > 0$ and a homeomorphism (continuous bijection with continuous inverse) $h : B_\varepsilon(x_0) \to \mathbb{R}^d$ from the ball of radius ε around x_0 to $\mathbb{R}^d$ mapping the connected components of the trajectories of F inside $B_\varepsilon(x_0)$ to the trajectories of the linearization $DF(x_0)$.

Example 9.6. Consider the undamped pendulum equation $f'' = -\alpha \sin(f)$ for $\alpha > 0$. Physically α is equal to g/ℓ where g is the gravitational constant and ℓ is the length of the pendulum.

As usual, we consider this as a first order equation in phase space

$$\mathbf{f}' = \begin{pmatrix} f' \\ f \end{pmatrix}' = \begin{pmatrix} -\alpha \sin(f) \\ f' \end{pmatrix} = F(\mathbf{f})$$

where

$$F \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -\alpha \sin(y) \\ x \end{pmatrix}.$$

This ODE has equilibrium points at $a_n = (0, n\pi)$ for every integer n. The derivative of F is

$$DF = \begin{pmatrix} \frac{dF_1}{\partial x} & \frac{dF_1}{\partial y} \\ \frac{dF_2}{\partial x} & \frac{dF_2}{\partial y} \end{pmatrix} = \begin{pmatrix} 0 & -\alpha \cos(y) \\ 1 & 0 \end{pmatrix},$$

so that at each equilibrium point a_n we have that

$$DF(a_n) = \begin{pmatrix} 0 & -\alpha \\ 1 & 0 \end{pmatrix}.$$

This matrix can be diagonalized

$$\begin{pmatrix} 0 & -\alpha \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} \sqrt{\alpha} & \sqrt{\alpha} \\ -i & i \end{pmatrix} \begin{pmatrix} i\sqrt{\alpha} & 0 \\ 0 & -i\sqrt{\alpha} \end{pmatrix} \begin{pmatrix} \sqrt{\alpha} & \sqrt{\alpha} \\ -i & i \end{pmatrix}^{-1},$$

so that the trajectories of the linearized ODE $(\mathbf{f} - a_n)' = DF(a_n)(\mathbf{f} - a_n)$ are ellipses around a_n of the form

$$\left\{ a_n + \begin{pmatrix} \sqrt{\alpha} & \sqrt{\alpha} \\ -i & i \end{pmatrix} \begin{pmatrix} e^{it\sqrt{\alpha}} & 0 \\ 0 & e^{-it\sqrt{\alpha}} \end{pmatrix} \begin{pmatrix} \sqrt{\alpha} & \sqrt{\alpha} \\ -i & i \end{pmatrix}^{-1} \begin{pmatrix} x \\ y \end{pmatrix} : t \in \mathbb{R} \right\}.$$

Since this fixed point is not hyperbolic, the linearization does *not* tell us whether or not the fixed point is stable. Let us focus on the fixed point at $(0, 0)$. The physical nature of the problem saves us by inspiring us to write down a Lyapunov function for the system: the energy! If we write down the function

$$E \begin{pmatrix} x \\ y \end{pmatrix} = \frac{1}{2}\ell^2 x^2 + g\ell(1 - \cos y)$$

where the first term represents the kinetic energy of the system and the second term represents the gravitational potential energy of the system, we can check by calculus that our solutions satisfy

$$\frac{d}{dt} E(\mathbf{f}) = 0.$$

In other words, energy is conserved. One can deduce from this that the trajectories of our ODE are exactly the level lines of the energy, and form "approximate ellipses" around $(0, 0)$. As such, in this example the linearization does give a good approximation to the true behaviour of the system around the equilibrium, even though this equilibrium is not hyperbolic.

Example 9.7. The ODEs governed by the vector fields

$$F\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} y + x(x^2 + y^2) \\ -x + y(x^2 + y^2) \end{pmatrix}$$

and

$$G\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} y - x(x^2 + y^2) \\ -x - y(x^2 + y^2) \end{pmatrix}$$

have the same linearization around the equilibrium point at $(0, 0)$, but the first is unstable and the second is asymptotically stable.

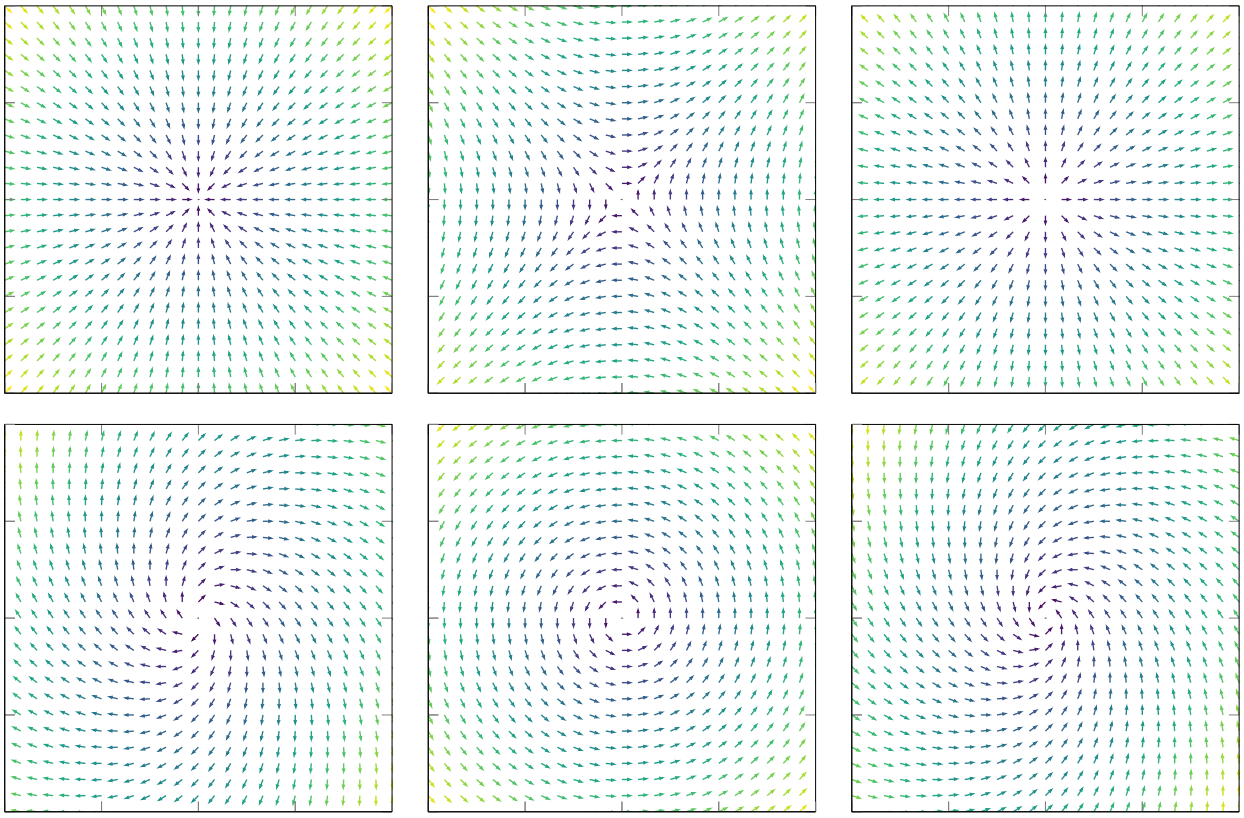


Figure 9: Unit vector field representations of six different linear autonomous systems in $\mathbb{R}^2$ that are diagonalizable with non-zero eigenvalues. Clockwise from top left: Sink, Saddle, Source, Spiral Sink, Center, Spiral Source.

Sink: Negative real eigenvalues. Example: $\mathbf{f}' = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \mathbf{f}$

Saddle: Real eigenvalues of different signs. Example: $\mathbf{f}' = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \mathbf{f}$

Source: Real positive eigenvalues. Example: $\mathbf{f}' = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \mathbf{f}$

Spiral Source: Complex eigenvalues, positive real parts. Example: $\mathbf{f}' = \begin{pmatrix} 1 & 1 \\ -1.5 & 1 \end{pmatrix} \mathbf{f}$

Center: Complex eigenvalues, zero real parts. Example: $\mathbf{f}' = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \mathbf{f}$

Spiral Sink: Complex eigenvalues, negative real parts. Example: $\mathbf{f}' = \begin{pmatrix} -1 & -1 \\ 1.5 & -1 \end{pmatrix} \mathbf{f}$

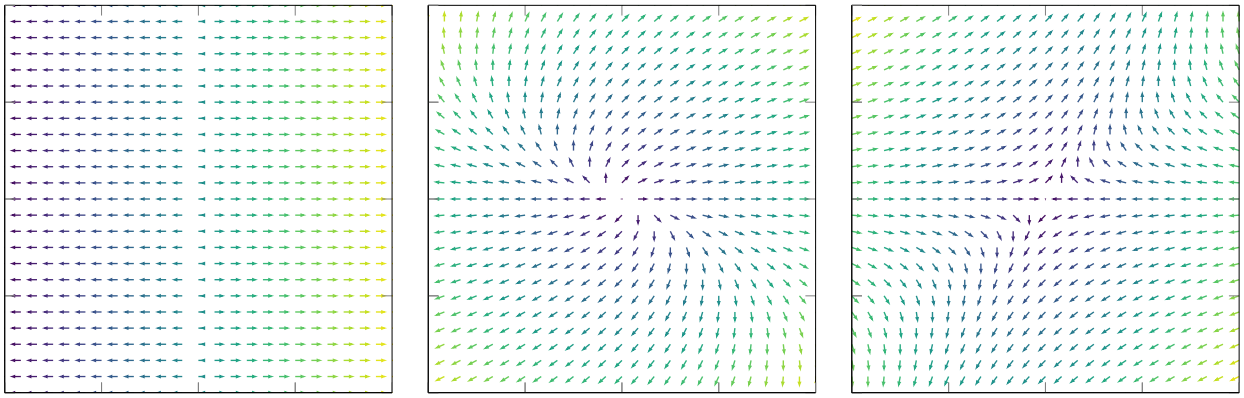


Figure 10: Left: A linear vector field with one zero eigenvalue and one negative real eigenvalue. Center: A non-diagonalizable vector field with positive real eigenvalue. Right: A non-diagonalizable vector field with negative real eigenvalue.