Notes for MAT 308, Spring 2025

Joseph Helfer

Last updated: May 8, 2025

Contents

1 Jan 27, Intro and calculus review

1.1 Introductory remarks

- Differential equations are the central formalism in all of the most important physical theories, and are of great importance in many other branches of science and engineering, as well as in pure mathematics.
  - For example: "Newton's Second Law", "Maxwell's Equations" from electromagnetism, "Einstein's Field Equations" from general relativity, "Schrödinger's equation" from quantum mechanics, are all systems of differential equations.
- Roughly speaking a differential equation is an equation containing (ordinary or partial) derivatives.
- Example: Newton's Second Law $F = ma$ is really $F = m\frac{\mathrm{d}^2 u}{\mathrm{d}t^2}$, where $u(t)$ is the position of a given object with mass m.
  - F typically depends on u and maybe on $\frac{\mathrm{d}u}{\mathrm{d}t}$.
  - For example, for a massive body in a central gravitation field, we have $F = -G\frac{u}{|u|^3}$ for some constant G, so the equation becomes $-G\frac{u}{|u|^3} = m\frac{\mathrm{d}^2 u}{\mathrm{d}t^2}$.
  - The "unknown" in this equation, for which one might try to solve, is u. It appears twice, with different amounts of derivatives. A solution would not be a *number* as in an ordinary equation but a *function* of time: the trajectory of the body in question.
- In Newton's equation, the variable u is actually a *vector*-valued – i.e., $\mathbb{R}^3$-valued – function.
  - Thus, the equation is actually a system of *three* equations.
  - As you learned in linear algebra, a system of linear equations is most fruitfully understood as a single matrix equation $Av = b$.
  - The same kind of thing is true in the theory of differential equations, and moreover there are some interesting and important further ideas in linear algebra which come in.
  - This is why linear algebra is also part of this course.
- Newton's equation only contains *ordinary* derivatives and is thus a so-called **Ordinary Differential Equation** (or **ODE**).
  - By contrast, all of the other equations mentioned above are **Partial Differential Equations** (or *PDE*s).
  - We will be dealing almost exclusively with ODEs in this class.
  - Just to give you an idea, a simple example of a PDE is the *heat equation*, which looks like this: $\frac{\partial u}{\partial t}(t,x) = -\frac{\partial^2 u}{\partial x^2}(t,x)$.

1.2 Calculus review

On engineers and mathematicians

- There are two ways of approaching the subject of differential equations.
  - The "mathematician's way" gives precise definitions of all the objects under consideration, and rigorous proofs on the basis of those definitions.

- The "engineer's (or physicist's) way" rests on a basic intuitive understanding of all of the basic objects (real numbers, continuous functions, limits, etc.) and their properties, and this intuition is based on the physical objects and phenomena that these mathematical objects represent.
- Feynman ("The character of physical law"): *There are two kinds of ways of looking at mathematics, which [...] I [...] call the Babylonian tradition and the Greek tradition. In Babylonian schools in mathematics the student would learn something by doing a large number of examples until he caught on to the general rule. [...] Also he would know a large amount of geometry, a lot of the properties of circles, the theorem of Pythagoras, formulae for the areas of cubes and triangles; in addition, some degree of argument was available to go from one thing to another. [...] But Euclid discovered that there was a way in which all of the theorems of geometry could be ordered from a set of axioms that were particularly simple. The Babylonian attitude – or what I call Babylonian mathematics – is that you know all of the various theorems and many of the connections in between, but you have never fully realized that it could all come up from a bunch of axioms.*

- **Both of these ways are valuable and important**.
  - And in fact, the "ideal mathematician" or "ideal engineer" don't exist; everyone is somewhere in between the two.
  - As this is a *mathematics* course (MAT 308), we must to some extent emphasize the "mathematician's way": in principle, you should know the precise definitions of all the objects we consider, be able to give precise statements of all the theorems we consider, and be able to follow the proofs, and reproduce the simpler ones. (In principle, this also means being familiar with the axioms of set theory – which, though a good thing, we will not insist on.)
  - Thus, we begin with a quick review of the basic elements of set theory and calculus which we will be using, to make sure we are all on the same page.

Sets

- We take for granted all the notions of set theory, and the basic properties of and operations with sets, which you know well by now.
- The most important sets are $\mathbb{N} = \mathbb{Z}_{\geq 0}, \mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$ for the naturals, integers, rationals, reals, and complex numbers, to all of which we shall return shortly.
  - We write $\mathbb{N}_+$ or $\mathbb{Z}_{>0}$ or something like that for the set of positive integers.
- The product of two sets is the set of all ordered pairs $X \times Y = \{(x, y) \mid x \in X, y \in Y\}$.
- We write $A \subset B$ or $A \subseteq B$ for "A is a subset of B".
  - If we want to express that A is a *proper* subset of B, we write $A \subsetneq B$
- A <u>relation</u> R between sets X and Y is a subset of the product $R \subset X \times Y$.
  - We write "xRy" for $(x, y) \in R$.
- A <u>function</u> (or <u>mapping</u> or <u>map</u>) $f\colon X \to Y$ is a relation $f \subset X \times Y$ such that for every $x \in X$, there is a unique $y \in Y$ with $xfy.$, i.e.

$$\forall x \in X, \exists! y \in Y, xfy.$$

  - We write $f(x)$ for the unique $y \in Y$ such that xfy.

- A(n infinite) sequence in a set X is a function $\mathbb{N} \to X$ (or $\mathbb{Z}_{\geq n} \to X$ for any $n \in \mathbb{Z}$).
  - As usual, we write $(x_i)_{i=0}^{\infty}$ for the sequence $\mathbb{N} \to X$ with $i \mapsto x_i$ for $i \in \mathbb{N}$.
- A finite sequence of length $n \in \mathbb{N}$ or n-tuple in a set X is a function $\{1, \ldots, n\} \to X$.
  - We write X^n for the set of finite sequences of length n.
  - As usual, we write $v = (v_1, \ldots, v_n)$ for the tuple $\{1, \ldots, n\} \to X$ given by $i \mapsto v_i$.
  - More generally, given sets $X_1, \ldots, X_n$, we write $X_1 \times \cdots \times X_n$ for the set of finite sequences $(x_1, \ldots, x_n)$ with $x_i \in X_i$ for each i.
- We write $X \cong Y$ if there exists a bijection between X and Y, and we write $f \colon X \xrightarrow{\sim} Y$ to indicate that f is a bijection.
- We write Y^X for the set of functions $X \to Y$ and $\mathcal{P}(X)$ for the power set of X, i.e., the set of subsets of X.
  - We have $2^X \cong \mathcal{P}(X)$ for any set X, where 2 is the set $\{0, 1\}$.

The real numbers

- We recall the main properties of the set of real numbers $\mathbb{R}$ with its operations $+ \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and $\cdot \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and its ordering " $\leq$ " $\subset \mathbb{R} \times \mathbb{R}$:
- **Field axioms**
  - addition and multiplication are commutative and associative, and multiplication distributes over addition
  - 0 and 1 are identity elements for addition and multiplication, respectively (i.e., $0 + a = a$ and $1 \cdot a = a$ for all $a \in \mathbb{R}$).
  - (*Exercise*: 0 and 1 are each uniquely determined by this property!)
  - Every element $a \in \mathbb{R}$ has an additive inverse $(-a)$ (i.e., $a + (-a) = 0$).
  - Every element $a \in \mathbb{R} - \{0\}$ has a multiplicative inverse a^{-1}.
- **Ordering axioms**
  - $\leq$ is reflexive, meaning $a \leq a$ for all $a \in \mathbb{R}$.
  - $\leq$ is antisymmetric, meaning $a \leq b \wedge b \leq a \Rightarrow a = b$ for all $a, b \in \mathbb{R}$.
  - $\leq$ is transitive, meaning $a \leq b \wedge b \leq c \Rightarrow a = c$ for all $a, b, c \in \mathbb{R}$.
  - $\leq$ is total, meaning $a \leq b \vee b \leq a$ for all $a, b \in \mathbb{R}$.
  - If $a \leq b$ then $a + c \leq b + c$ for any $a, b, c \in K$.
  - If $a \leq b$ and $c \geq 0$, then $a \cdot c \leq b \cdot c$ for any $a, b, c \in K$.
- **Completeness axiom**
  - Given a set $S \subset \mathbb{R}$, we say that $a \in \mathbb{R}$ is an upper bound for S if $s \leq a$ for all $s \in \mathbb{R}$, and we say that S is bounded-above if it has some upper bound.
  - Completeness axiom: each bounded-above subset $S \subset \mathbb{R}$ has a least upper bound (or supremum) $\sup S$, meaning that $\sup S$ is an upper bound for S, and that $\sup S \leq a$ for every upper bound a for S.
- In short, one says that $(\mathbb{R}, +, \cdot, \leq)$ is a complete ordered field.

- On the basis of these axioms, one can prove all the familiar properties of the algebraic operations and the ordering relation (an activity you have hopefully done before, and should try your hand at if you haven't).
  - In fact, these axioms **completely determine** the real numbers in the following sense: given any system $(K, \tilde{+}, \tilde{\cdot}, \tilde{\leq})$ consisting of a set K, binary operations $\tilde{+}$ and $\tilde{\cdot}$ on K, and a binary relation $\tilde{\leq}$ on K satisfying the above axioms, there exists a unique bijection $F\colon \mathbb{R} \to K$ satisfying $F(a+b) = F(a)\tilde{+}F(b)$, $F(a \cdot b) = F(a)\tilde{\cdot}F(b)$, and $a \leq b \iff F(a)\tilde{\leq}F(b)$ for all $a, b \in K$.
    - (Such a bijection F is called an "isomorphism of ordered fields".)
    - (The aforementioned theorem establishes the *uniqueness* (up to isomorphism) of the real numbers. One may still reasonably inquire about their *existence*: why must there exist such a set satisfying these axioms at all? We simply take it for granted that it exists; if one wants to *construct* it, this must be on the basis of some more basic axioms: either the Peano axioms of arithmetic, or else the axioms of set theory.)
- The other number systems
  - We define the sets $\mathbb{N}, \mathbb{Z}, \mathbb{Q} \subset \mathbb{R}$ as follows.
  - $\mathbb{N} \subset \mathbb{R}$ is the smallest subset such that $0 \in \mathbb{N}$ and such that $n+1 \in \mathbb{N}$ whenever $n \in \mathbb{N}$.
    - Concretely, this means that $\mathbb{N}$ is the intersection $\bigcap_{S \subset \mathbb{R}} S$ of all subsets of $\mathbb{R}$ satisfying these two properties.
    - It follows immediately from this that $\mathbb{N}$ satisfies the all-important **principle of mathematical induction**: given any $S \subset \mathbb{N}$ such that $0 \in S$ and $n+1 \in S$ whenever $n \in S$, we have $S = \mathbb{N}$.
  - $\mathbb{Z} = \{a - b \mid a, b \in \mathbb{N}\} \subset \mathbb{R}$.
  - $\mathbb{Q} = \{a/b \mid a, b \in \mathbb{Z} \wedge b \neq 0\} \subset \mathbb{R}$.
  - Finally, $\mathbb{C}$ is simply defined to be $\mathbb{R}^2$.
    - Given $(a_1, b_1), (a_2, b_2) \in \mathbb{C}$, we set $(a_1, b_1) + (a_2, b_2) := (a_1 + b_1, a_2 + b_2)$, and $(a_1, b_1) \cdot (a_2, b_2) := (a_1 a_2 - b_1 b_2, a_1 b_2 + a_2 b_1)$.
    - One may check that this makes $\mathbb{C}$ into a *field* (though of course, not an *ordered* field, since there is no natural ordering of points in the plane).
    - Given $a \in \mathbb{R}$, we write a as a shorthand for $(a, 0) \in \mathbb{C}$.
    - We also write i as a shorthand for $(0, 1) \in \mathbb{C}$; we have $i^2 = -1$.
    - Thus, we may write $(a, b) = a + bi$ for every $(a, b) \in \mathbb{C}$, and we then have $(a_1 + b_1 i) + (a_2 + b_2 i) = (a_1 + b_1) + (a_2 + b_2)i$ and $(a_1 + b_1 i) \cdot (a_2 + b_2 i) = a_1 a_2 + b_1 b_2 i^2 + a_1 b_2 i + a_2 b_1 i$.

2 Jan 29, More calculus review and 10.1: Direction fields

2.1 Calculus review continued

Calculus

- We will be reviewing various topics from calculus (and linear algebra) as they come up; we will just recall a couple of the most important ones here
- We consider functions $f\colon I \to \mathbb{R}$ defined on some domain $I \subset \mathbb{R}$.
  - Typically, $I \subset \mathbb{R}$ will be an interval: a set of the form (a, b), $[a, b)$, $(a, b]$, or $[a, b]$, where $a \in \mathbb{R} \cup \{-\infty\}$ and $b \in \mathbb{R} \cup \{\infty\}$ and $a \leq b$.
- Limits:
  - Given a function $f\colon I \to \mathbb{R}$ and $x \in I$, we have the limit $\lim_{x \to a} f(x)$ and the one-sided limits $\lim_{x \nearrow a} f(x)$ and $\lim_{x \searrow a} f(x)$ (each of which may or may not exist).
  - Ideally, you should know the **definition** of the limit: $\lim_{x \to a} f(x) = b$ means $\forall \epsilon > 0, \exists \delta > 0, \forall x \in I, (0 < |x - a| < \delta \Rightarrow |f(x) - b| < \varepsilon)$.
  - But the most important thing is that you know the basic **rules** for limits: $\lim_{x \to a} f(x) + g(x) = \lim_{x \to a} f(x) + \lim_{x \to a} g(x)$ (assuming the right-hand sides exists), and so on.
- A function $f\colon I \to \mathbb{R}$ is continuous at $a \in I$ if $\lim_{x \to a} f(x) = f(a)$, and is continuous on I if it is continuous at each $a \in I$.
  - Continuous functions satisfy the **intermediate value theorem**.
- Derivatives:
  - $f\colon I \to \mathbb{R}$ is differentiable at $a \in I$ if $\lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$ exists, and if so, this limit is called $f'(a)$.
  - f is differentiable on I if it is differentiable at each $a \in I$; in this case, we have the derivative $f'\colon I \to \mathbb{R}$.
    - **Theorem:** Differentiability implies continuity.
  - We sometimes write $y = f(x)$ and then denote the derivative by $f'(x) = \frac{\mathrm{d}y}{\mathrm{d}x}$.
    - Here, we are thinking of x and y as "variables", with y varying "dependently" on x; then $\mathrm{d}x$ represents an "infinitesimal variation" of x, $\mathrm{d}y$ represents the resulting infinitesimal variation of y, and the derivative is their quotient.
    - This is the classical way of thinking about functions, which has been preserved by the physicists and engineers, but among mathematicians has been replaced by the set-theoretic perspective.
    - With some care, this notation can be used in a consistent and rigorous way (in fact, doing so leads to some very interesting mathematics), but in general, it is best to follow the physicists and engineers, and freely use our intuition about infinitesimals without worrying about the formal definition; and then if we want a rigorous proof, we can always revert to the formal, set-theoretic definitions.
    - In any case, we often simply write $\frac{\mathrm{d}f}{\mathrm{d}x}$ for f'.
  - We write f'' or $\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}$ for the derivative of f' (if it exists), and f''' and $f^{(4)}$ and $f^{(5)}$ and so on.
  - Again the most important thing is to know all the **rules** for derivatives: the sum rule, product rule, quotient rule, **chain rule**, the rule for x^n, and the derivatives of everyone's favourite functions: sin, cos, e^x, $\ln x$, and so on.

- We denote by $\mathcal{C}^k(I)$ the set of functions $f\colon I \to \mathbb{R}$ such that the derivatives $f', f'', \ldots, f^{(k)}$ exist and are continuous; $\mathcal{C}^0(I)$ is simply the set of continuous functions on I, and $\mathcal{C}^\infty(I)$ is the set of infinitely differentiable (or smooth) functions on I. All of the most common functions are smooth.

- Partial derivatives
  - We consider functions $f\colon U \to \mathbb{R}$ defined on a domain $U \subset \mathbb{R}^n$.
  - Often, U will be all of $\mathbb{R}^n$, or maybe some product of intervals $U = I_1 \times \cdots \times I_n$.
  - The definition of **limit** and **continuity** for such functions is exactly the same as in the single-variable case,
    - except that now, the absolute values $|x - a|$ and so on appearing in the definition are now considered norms.
    - We recall the norm of $v \in \mathbb{R}^n$ is given as in the Pythagorean theorem: $|v| = \left(\sum_{i=1}^n v_i^2\right)^{1/2}$.
  - Given $f\colon U \to \mathbb{R}$ and $a \in U$, the i-th partial derivative of f at a, if it exists, is the "derivative of f at a in the i-th coordinate direction, holding all other coordinates constant".
  - For example, if $n = 2$, the two partial derivatives at $(a, b) \in U$ are
$$\lim_{h\to 0} \frac{f(a+h, b) - f(a, b)}{h} \qquad \text{and} \qquad \lim_{h\to 0} \frac{f(a, b+h) - f(a, b)}{h}.$$
  - In general, the i-th partial derivative at $a \in U$ is
$$\lim_{h\to 0} \frac{f(a_1, \ldots, a_{i-1}, a_i + h, a_{i+1}, \ldots, a_n) - f(a_1, \ldots, a_n)}{h}.$$
  - There are many notations for the i-th partial derivative; the simplest are $\partial_i f$ or f_i; often we write $y = f(x_1, \ldots, x_n)$, and then write $\frac{\partial y}{\partial x_i}$ or $\frac{\partial f}{\partial x_i}$ or $\partial_{x_i} f$ or f_{x_i}.
  - In practice, one evaluates the partial derivative with respect to x_i by "pretending all the other variables are constant and taking the ordinary derivative with respect to x_i".

- The chain rule
  - The sum, product, quotient rules for partial derivatives are just like those for ordinary derivatives.
  - The **chain rule** is more interesting: if $y_i = f_i(x_1, \ldots, x_m)$ for $i = 1, \ldots, n$ and $z = g(y_1, \ldots, y_n)$, so that
$$z = g\big(f_1(x_1, \ldots, x_m), \ldots, f_n(x_1, \ldots, x_m)\big),$$
    then
$$\frac{\partial z}{\partial x_i} = \sum_{j=1}^n \frac{\partial z}{\partial y_j} \cdot \frac{\partial y_j}{\partial x_i}.$$
  - The chain rule is most elegantly expressed using the **Jacobian** (or **total derivative**) matrix. For this, we must recall a bit about vector-valued functions and matrix multiplication, and we will do this when it comes up.

- Sequences and series
  - The limit $\lim_{n\to\infty} x_n$ of a sequence $(x_n)_{n=0}^\infty$, if it exists, is defined as the unique $a \in \mathbb{R}$ such that $\forall \epsilon > 0, \exists N \in \mathbb{N}, \forall n \geq N, |x_n - a| < \varepsilon$.
    - We also write $x_n \xrightarrow{n\to\infty} a$.

- Again, the most important thing is that you know the main **rules** for limits: the sum rule $\lim_{n\to\infty} a_n + b_n = \lim_{n\to\infty} a_n + \lim_{n\to\infty} b_n$ (assuming the right-hand side exists), the difference and quotient rules, and the rule that if f is a function which is continuous at a and $x_n \xrightarrow{n\to\infty} a$, then $f(x_n) \xrightarrow{n\to\infty} f(a)$.
- The sum of the infinite series $\sum_{n=0}^{\infty} a_n$, if it exists, is the limit $\lim_{N\to\infty} S_n$ of the partial sums $S_N = \sum_{n=0}^{N} a_n$.
- You should know the various **tests** for convergence: the comparison test, alternating series test, absolute convergence test, ratio test, root test, integral test.

- Integration
  - If $I \subset \mathbb{R}$ contains the interval $[a, b]$ and $f\colon I \to \mathbb{R}$ is any function, we can talk about the integral $\int_a^b f(x)\,\mathrm{d}x$, which may or may not exist, but always *does* exist when f is continuous, or even continuous outside of finitely many points.
  - Again, it is best – though perhaps not essential – to know the **definition** of the integral, in terms of Riemann sums: at least for continuous functions, the integral $\int_a^b f(x)\,\mathrm{d}x$ is given by the limit $\lim_{N\to\infty} \frac{b-a}{N} \sum_{i=1}^{N} f(a + i\frac{b-a}{N})$.
    - (In general (for not necessarily continuous f), the definition is a bit more complicated: one has to consider arbitrary partitions $a = x_0 \leq \cdots \leq x_N = b$, and not just the "regular partition" $x_i = a + i\frac{b-a}{N}$.)
    - The notation $\int_a^b f(x)\,\mathrm{d}x$ is meant to represent an "infinite sum of infinitesimals" (the integral sign is an "S" for "sum"). This is the classical notion, which again has been superseded by the modern set-theoretic definition. Again, it can be made rigorous with some effort, and again, we do well to follow the physicists and engineers in freely using this intuition when it is useful.
  - Again, the most important thing to know are the various **rules**, among which the most important is the **fundamental theorem of calculus**:
    - The integral $\int_a^b f(x)\,\mathrm{d}x$ is equal to $F(b) - F(a)$ where $F\colon [a, b] \to \mathbb{R}$ is any anti-derivative of f, meaning a function with $F' = f$ (assuming that an anti-derivative exists!).
    - Moreover, if f is continuous, then an anti-derivative *does* exist.
    - When anti-derivative F of f exists is uniquely determined up to a constant $C \in \mathbb{R}$, and the "indefinite integral" is defined as $\int f(x)\,\mathrm{d}x = F(x) + C$.
  - Other rules you should know include: the sum and difference rule, the substitution rule, and integration by parts, as well as the million little tricks you've learned to apply these in various special cases.
  - We will review the important topics of multi-variable integrals and improper integrals if and when they come up.

2.2 10: First-order ODEs

- We now begin our study of differential equations.
  - As a preliminary definition, a *differential equation of order n* is an equation whose unknown is a function $y(x)$, and which may involve the derivatives of y up to order n, that is $y, y', y'' \ldots, y^{(n)}$, as well as x itself.
  - Examples of order 1 and 2 are

$$y'(x) + y(x) = x \qquad \text{and} \qquad y''(x) + y'(x) = 0.$$

which we usually just write as

$$y' + y = x \qquad \text{and} \qquad y'' + y' = 0.$$

- A solution to this equation on an interval $I \subset \mathbb{R}$ is then a function $y\colon I \to \mathbb{R}$ which satisfies this equation for all $x \in I$ (in particular, y must be n times differentiable).
  - In principle, one can consider differential equations on *any* domain $I \subset \mathbb{R}$, and not just on an interval I. But if I is not connected, say for example, $I = I_1 \cup I_2$, where I_1 and I_2 are disjoint intervals, then a solution to the differential equation on I just amounts to a solution on I_1 and a separate solution on I_2.
- For example, $y(x) = x - 1$ and $y(x) = e^{-x}$ are solutions to the above two equations.
- (As with ordinary equations, we are often not interested in finding *a* solution, but in finding *all* possible solutions.)

- The above definition will be perfectly satisfactory for our purposes, though it is also good to give a more precise, formal definition of what we mean by a "differential equation", which one can do as follows.
  - We define a(n ordinary) differential equation of order n to be an arbitrary function $F\colon \mathbb{R}^{n+1} \to \mathbb{R}$ (or more generally $U \to \mathbb{R}$ for some $U \subset \mathbb{R}^{n+2}$).
    - Intuitively, we think of F as representing the equation $y^{(n)} = F(x, y, y', \dots, y^{(n-1)})$.
    - (One can more generally consider so-called "implicit" equations given by a function $E(x, y, y', \dots, y^{(n)})$, which we think of as representing the equation $0 = F(x, y, y', \dots, y^{(n)})$, but this is rarely done.)
  - A solution to the differential equation F on $I \subset \mathbb{R}$ is then a n-times differentiable function $y\colon I \to \mathbb{R}$ such that $((x, y(x), y'(x), \dots, y^{(n-1)}(x)) \in U$ and) $y^{(n)}(x) = F(x, y(x), y'(x), \dots, y^{(n-1)}(x)) = 0$ for all $x \in I$.
  - For example, in this formalism, the above two equations would be given by the functions $F(x, y_0) = x - y_0$ and $E(x, y_0, y_1) = -y_1$, respectively.
  - An example where we have to take some $U \subsetneq \mathbb{R}^n$ is the equation $y'' = 1/y' + y$, where $F(x, y_0, y_1) = 1/y_1 + y_0$, and so the largest U we can take is $\{(x, y_0, y_1) \in \mathbb{R}^4 \mid y_1 \neq 0\}$.

- To begin with, we will be considering only **first-order** differential equations.

2.3 10.1: Direction fields

- We consider an arbitrary first-order equation

$$y' = F(x, y).$$

  - Any solution $y(x)$ passing through (x_0, y_0) must have slope $F(x_0, y_0)$ at that point.
  - Thus, we think of $F(x, y)$ as assigning a slope to each point (which we can represent as a short line segment through that point with that slope); this is called a **slope field** or **direction field**.
  - This is not to be confused with a **vector field**.
  - We can draw the direction field $F(x, y)$ by sampling several points and drawing line segments, or we can first find several solutions to the differential equation, and then simply plot various tangents along the resulting graphs.

Example 10.1.1

- Consider the equation $y' = -y/x$ for $x \neq 0$.
- We sample the slopes at a few points:

(x, y)	$y' = -y/x$
(1,1)	-1
(1,2)	-2
(2,1)	-1/2
(-1,2)	2
(-2,2)	1

- We obtain a picture as in Figure 10.2 (a)
- Let's solve the equation!
  - Let $I \subset \mathbb{R}$ be one of $(-\infty, 0)$ or $(0, \infty)$.
  - Suppose that $y\colon I \to \mathbb{R}$ is a solution to $y' = -y/x$.
  - Rearrange: $xy' + y = 0$ (equivalent by the assumption $x \neq 0$).
  - Product rule: $(xy)' = 0$.
  - Thus $x \cdot y(x)$ is some constant c.
  - So $y = c/x$.
- Thus, every solution is of the form $y = c/x$.
  - Conversely, since each step above was a **logical equivalence**, it follows that $y = c/x$ **is** always a solution (as one can also check directly).
  - We may say that the general solution to the equation is $y = c/x$ with $x \in \mathbb{R}$.
- We can plot the solutions and see they are tangent to our slope field.

3 Feb 3, 10.2: Separation of variables

3.1 10.1: Direction fields

Example 10.1.2

- If $F(x, y)$ is independent of y, the equation takes the form

$$y' = G(x);$$

  assume G is continuous.

- Then the solutions are simply the antiderivatives of G: $y = \int G(x)\,\mathrm{d}x + C$.
  - All of the solution curves will be parallel to each other, as one can also see from the direction field.
- Example: $y' = \cos x$.
  - Then the general solution is $y = \sin x + c$.
  - Note: when discussing the solutions to a differential equation, one should always first state *on which interval* I one is seeking solutions $y\colon I \to \mathbb{R}$, though this is often left implicit.
  - So here, we should say we are seeking solutions y defined on all of $\mathbb{R}$, and then the general such solution is $y = \sin x + c$.

Example 10.1.3

- We see that differential equations tend to have many solutions.
- We can add extra conditions to a differential equation to single out a particular solution.
- If we demand that y passes through a given point (x_0, y_0), i.e., that $y(x_0) = y_0$, this is called an **initial condition**. (More generally, we can demand that $y^{(i)}(x_i) = y_i$.)
- The problem of satisfying a differential equation with a given initial condition is called an **initial value problem** (IVP).
- The name comes from thinking of x as the "time parameter" and x_0 as the "initial time".
- We return to $y' = -y/x$, $x \neq 0$ with solutions $y = c/x$.
  - Now consider the initial value problem $(x_0, y_0) = (1/2, 2)$.
  - We thus have $2 = c/(1/2)$ and hence $c = 1$.
  - The (unique!) solution to the IVP is thus $y = 1/x$, $x > 0$.
- We can also solve the general IVP for a given (x_0, y_0).
  - We have $y_0 = c/x_0$, so $c = x_0 y_0$, and the solution is $y = x_0 y_0/x$, with $x \in (0, \pm\infty)$ depending on the sign of x_0.

3.2 10.2B: Separation of variables

- Among the first-order equations $y' = F(x, y)$, we have seen that the simplest case is when $F(x, y)$ doesn't depend on y. We can then solve it simply by integration.
- We now study the more general case in which in which $F(x, y)$ is of the form $f(x) \cdot h(y)$.
  - Supposing h is non-zero and setting $g(y) = h^{-1}(x)$, these are thus the equations of the form $g(y) \cdot y' = f(x)$.
  - These can often also be solved simply by integration, using a method called "separation of variables".
  - Writing $g(y)\frac{\mathrm{d}y}{\mathrm{d}x} = f(x)$, we obtain "$g(y)\,\mathrm{d}y = f(x)\,\mathrm{d}x$" and hence
$$\int g(y)\,\mathrm{d}y = \int f(x)\,\mathrm{d}x.$$
  - Computing antiderivatives F and G of f and g, we thus have $G(y) = F(x) + C$, and can thus find y as long as we can solve for y in this (non-differential-)equation.
  - (We can do the above computation without using "differentials": we simply have $\frac{\mathrm{d}G(y(x))}{\mathrm{d}x} = y'(x) \cdot g(y(x)) = f(x)$ by the chain rule, and hence $G(y(x))$ is an antiderivative of $f(x)$.)

Example 10.2.3

- We consider the differential equation $\frac{\mathrm{d}P}{\mathrm{d}t} = kP$, where $k > 0$ is a constant.
  - This describes the rate of growth of a population of bacteria of size $P(t)$ (say in grams), in which each bacterium reproduces at a rate of k bacteria per second.
  - As always in mathematical modelling, we are making simplifying assumptions in this equation, so we can only expect the solution of the equation to give an approximation to the population growth of the bacteria.
- We know from experience that the solution is $P(t) = Ke^{kt}$ for some constant K.
  - But let us see how we could have arrived at that using separation of variables.
  - Assuming P is always positive, we have $\int \frac{1}{P}\,\mathrm{d}P = \int k\,\mathrm{d}t$, hence $\ln P = kt + C$, hence $P(t) = e^{kt+C} = e^C \cdot e^{kt}$.
  - The solution to the IVP $P' = kP; P(0) = P_0$ is thus $P(t) = P_0 \cdot e^{kt}$.
  - (Assuming instead that P is always *negative* would given $\ln(-P) = kt + C$ and hence $P(t) = -e^C \cdot e^{kt}$ and thus again $P(t) = P_0 e^{kt}$; assuming just that P is non-zero would also lead to the same conclusion, since being differentiable and hence continuous, it is then always positive or negative.)
  - **Be careful**: in this computation, we had to make the assumption that P is non-zero, so that we could divide by it. If this assumption wasn't valid, our solution may be incorrect. However, we can just check directly that our solution *is* correct.
- This obviously isn't a realistic solution for large t; one problem is that our model doesn't include the amount of *food* available to the bacteria.
- If we look carefully at what we did, we see that we *almost* showed that this is the unique solution to the IVP; the only problem is we had to assume P was non-zero; the possibility remains that there are other solutions which are zero somewhere (besides the obvious one $P(t) = 0$).

- The trick to show uniqueness without this assumption is to consider the product $P(t) \cdot e^{-kt}$.
- We then have $\frac{\mathrm{d}}{\mathrm{d}t}(Pe^{-kt}) = kPe^{-kt} - kPe^{-kt} = 0$. Hence $e^{-kt} \cdot P(t)$ is equal to a constant K and $P = Ke^{-t}$, as desired.
- We will return to this trick of "exponential multipliers" soon.
- (We also have uniqueness by the general uniqueness theorem we stated, but that is overkill.)

Example 10.2.6

- We consider a tank of a chemical solution, in which a particular chemical is flowing in and out at a particular rate.
- We let $S = S(t)$ be the amount of chemical at time t. Then

$$\frac{\mathrm{d}S}{\mathrm{d}t} = (\text{rate of inflow}) - (\text{rate of outflow}).$$

- Example: a 100-gallon tank contains 150 pounds of salt of salt in solution at time $t = 0$.
  - A salt solution with 2 pounds of salt per gallon is being added at a rate of 2 gallons per minute.
  - The solution, which we take to be homogeneous, is flowing out of the tank at a rate of 2 gallons per minute.
  - Thus salt is flowing in at 4 pounds per minute, and out at $2S(t)/100$ pounds per minute at time t.
- Thus S satisfies the IVP $\frac{\mathrm{d}S}{\mathrm{d}t} = 4 - \frac{2S}{100}; S(0) = 150$.
- We solve it using separation of variables
  - We have $\int \frac{\mathrm{d}S}{S-200} = -\int \frac{\mathrm{d}t}{50}$; we see that this will only be valid if $S - 200$ remains non-zero for all time. Since $S(0) - 200 = -50$, we suppose it remains negative for all t.
  - Hence $\ln\big(-(S-200)\big) = -t/50 + C$.
  - Hence $S = 200 - e^C e^{-t/50}$.
  - The initial condition gives $150 = S(0) = 200 - e^C$ and hence $e^C = 50$.
  - Hence $S = 200 - 50e^{-t/50}$; this is plotted in Figure 10.9; it has a positive slope and an asymptote $\lim_{t\to\infty} S = 200$.
  - Checking directly, we see that this is indeed a solution to the equation.
- (Again, we had to assume $S \neq 200$ here, so we haven't proven that this is the unique solution; again, we can circumvent this using an exponential multiplier.)
  - Note however that the uniqueness *is* important, from a scientific perspective: to draw the conclusion that the amount of chemical in the tank evolves in the way that we found, on the basis of the fact that it solves the given IVP, we have to know that it is the **only** solution to the IVP.

4 Feb 5, 10.3: More separation of variables; existence and uniqueness theorem

4.1 More of 10.2B: Separation of variables

Example 10.2.4

- Consider $y' = y/x$, for $x \in \mathbb{R} - \{0\}$.
  - The direction field is shown in Figure 10.7.
- By separation of variables we have $y^{-1}\,\mathrm{d}y = x^{-1}\,\mathrm{d}x$, hence $\ln|y| = \ln|x| + C$, hence $|y| = e^C|x|$, hence $y = \pm e^C x$.
  - Thus, the solutions are $y = kx$ for $k \in \mathbb{R}$.
  - (We should really be more careful here and say that we are looking for solutions on $(-\infty, 0)$ or on $(0, \infty)$, since otherwise we could also have solutions like
$$y(x) = \begin{cases} x & x < 0 \\ 2x & x > 0. \end{cases}$$
  - (The same is true in general when we say $\int x^{-1}\,\mathrm{d}x = \ln|x| + C$: if we take the function x^{-1} to be defined on all of $\mathbb{R} - \{0\}$, then the anti-derivative is only determined up to *two* constants: one on $(-\infty, 0)$ and one on $(0, \infty)$.)
- To prove uniqueness without needing y to be non-zero, we can use a similar trick to last time: we have $0 = y' - y/x$ and hence $0 = y'/x - y/x^2 = \frac{\mathrm{d}(y/x)}{\mathrm{d}x}$. Hence y/x is a constant.
  - (This is secretly again a case of the method of "exponential multipliers".)

4.2 End of 10.1: existence and uniqueness

Example 10.1.4

- In the previous examples, there was a unique solution to each IVP $y' = F(x, y); y(x_0) = y_0$, and this solution extended over the whole given domain; both of these features can fail.
- Consider
$$y' = \begin{cases} \sqrt{y}, & y \geq 0 \\ 0, & y < 0. \end{cases}$$
  - The direction field for this is shown in Figure 10.3 (a).
  - It has two solutions $y\colon \mathbb{R} \to \mathbb{R}$ passing through $(x_0, y_0) = (0, 0)$: $y(x) = 0$ and
$$y(x) = \begin{cases} 0, & x < 0 \\ x^2/4, & x \geq 0. \end{cases}$$
    - To conclude that y is differentiable and compute $y'(0)$, we use the **Theorem** from calculus that if a function f has a "left" derivative $\lim_{h \nearrow 0}(f(a+h)-f(a))/h$ and a "right" derivative $\lim_{h \searrow 0}(f(a+h)-f(a))/h$ at a given point a, and they agree, then f is differentiable at a.
  - In fact, there are infinitely many solutions to this IVP.
- As we will see, there is a simple condition on $F(x, y)$ which guarantees that this cannot not happen.

Example 10.1.5

- Now consider $y' = 1 + y^2$.
  - The direction field is shown in Figure 10.3 (b).
  - It has the solution $y = \tan x$ passing through $(x_0, y_0) = (0, 0)$.
  - This can only be extended to $x \in (-\pi/2, \pi/2)$ since it tends to $\pm\infty$, despite $F(x, y)$ being well-behaved everywhere.
  - In particular, had we stated the problem in the form "find a solution to this IVP which is defined on all of $\mathbb{R}$", the conclusion would have been that *this problem does not have a solution.*
- Again, this can be avoided by putting a simple condition on F.

Existence and uniqueness theorem

- We now state a general theorem which guarantees that a unique solution to a given IVP exists, if we assume that F satisfies certain conditions which in particular rule out the above pathologies.
  - The existence result should be compared to analogous phenomena for polynomial equations.
  - For degree 2 polynomial equations, we have an explicit method of solution, and even an explicit formula; this is analogous to the explicit methods we have introduced and will introduce to solve certain classes of differential equations.
  - On the other hand, we also know the general result (whose proof is easy): *every odd degree polynomial equation has a solution.* In this case, we are guaranteed on general grounds that a given equation has a solution, but we are not given any means to find it. The following existence theorem is of a similar nature.
- **Theorem**: Suppose $F\colon U \to \mathbb{R}$ is continuous, where $U = I_1 \times I_2$ for some intervals $I_1, I_2 \subset \mathbb{R}$, and that $F_y\colon U \to \mathbb{R}$ exists and is continuous.
  - Then for any $(x_0, y_0) \in U$, the IVP $y' = F(x, y); y(x_0) = y_0$ has a solution defined on some interval $I \subset I_1$, and this solution is unique in the sense that for any two solutions $y_1\colon I \to \mathbb{R}$ and $y_2\colon I' \to \mathbb{R}$, $y_1(x) = y_2(x)$ for $x \in I \cap I'$.
  - If moreover $I_2 = \mathbb{R}$ and there exists $B \in \mathbb{R}$ with $F_y(x, y) < B$ for all $(x, y) \in U$, then the solution will exist on the entire interval I_1.
- The first condition fails for Example 10.1.4, since $y \mapsto \sqrt{y}$ is not differentiable at $y = 0$.
  - The second condition fails for Example 10.1.5, since $1 + y^2$ is not bounded as $y \to \pm\infty$.

5 Feb 10, Proof of existence and uniqueness; 10.1: Numerical methods; and 10.3: Linear equations

5.1 End of 10.1: existence and uniqueness

- We will sketch the proof of the existence and uniqueness statement from last time.
- The proof of the second part is fairly easy: the point is that if the slope of $y(x)$ is bounded, then it can only increase by a finite amount in finite time by the **mean value theorem**.
  - The general fact is that if a function $y\colon I \to \mathbb{R}$ satisfies $|y'(x)| < B$ for all $x \in I$, then for any $a < b$ in I, we have $|y(b) - y(a)| < B(b-a)$.
  - Indeed, if we had $|y(b) - y(a)|/(b-a) \geq B$, then the mean value theorem would given $y'(\xi) = |y(b) - y(a)|/(b-a) \geq B$ for some $\xi \in (a,b)$, contradicting the assumption that $y'(x) < B$ for all $x \in I$.
  - One then has to show that the only way that a solution to the IVP can *fail* to extend over the entire interval I is if it diverges to $\pm\infty$.
- The proof of the first part uses an ingenious trick called **Picard iteration**.
  - One first converts the equation to the equivalent **integral equation** $y(x) = y_0 + \int_{x_0}^{x} F(x, y(x))\,\mathrm{d}x$; this integral exists since F is continuous.
  - One then finds a sequence of better and better **approximations** to the solution:
  - The first is simply $f_1(x) = y_0$.
  - Then we inductively define $f_{n+1}(x) = y_0 + \int_{x_0}^{x} F(x, f_n(x))\,\mathrm{d}x$.
  - Finally, we define $y(x) = \lim_{n\to\infty} f_n(x)$.
  - We then have

    $$y(x) = \lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} f_{n+1}(x) = \lim_{n\to\infty} y_0 + \int_{x_0}^{x} F(x, f_n(x))\,\mathrm{d}x = y_0 + \int_{x_0}^{x} F(x, y(x))\,\mathrm{d}x,$$

    as desired.
  - The tricky part is to show that the limit defining y **actually converges** (and that the exchanging of limit and integral in the last equation is legitimate); this is where the assumption is used that F_y exists and is continuous.
- (This is an instance of the general technique of *finding fixed points using iteration*: given a domain U and a continuous function $G\colon U \to U$, if we want to find a *fixed point* of G, i.e., a point $x \in U$ with $G(x) = x$, we can choose some arbitrary $x_0 \in U$, and iteratively define $x_{n+1} = G(x_n)$, and set $x = \lim_{n\to\infty} x_n$ *if this limit exists.* Using the continuity of G, we then have $G(x) = G(\lim_{n\to\infty} x_n) = \lim_{n\to\infty} G(x_n) = \lim_{n\to\infty} x_{n+1} = x$, as desired.)

5.2 10.1B: Numerical Methods

- In scientific applications, it is often important not to explicitly *solve* a given IVP $y' = F(x,y); y(x_0) = y_0$ (which often cannot be done anyway), but to compute numerically an approximation to its solution.
  - This amounts to running a simulation of a quantity y whose rate of change at each time x is given by $F(x,y)$.
  - Geometrically, it amounts to tracing out a curve tangent to a given direction field.

- A straightforward way to do this **Euler's method**:
  - Fix some step size $h > 0$.
  - We are given x_0 and y_0.
  - Now define $x_n = x_0 + n \cdot h$ for $n > 0$, and recursively define $y_{n+1} = y_n + F(x_n, y_n) \cdot h$.
  - It is easy to implement this in any programming language; an example is given in the book.
- There are many ways to improve this algorithm, and there is a whole field dedicated to studying such things.
  - A central problem that arises is the accumulation of **rounding errors**, for example if one tries to make the above step-size parameter h too small.
  - One such improved algorithm is discussed in the book.

5.3 10.3: Linear equations

- A first-order equation $y' = F(x, y)$ is called linear if F is of the form $F(x, y) = -g(x)y + f(x)$.
  - The equation can then be written as $y' + g(x)y = f(x)$; this is called its **normalized form**.
  - The name *linear* corresponds to the fact that $L(y) = y' + g(x)y$ is a *linear* function of y: we have $L(y_1 + y_2) = L(y_1) + L(y_2)$ and $L(a \cdot y) = a \cdot L(y)$ for $a \in \mathbb{R}$ (we will return to this when we do some linear algebra review).
  - Thus, a linear equation has the form $L(y) = f(x)$ for some linear operator L.
  - Just like in linear algebra, we call the equation homogeneous if $f(x) = 0$ and inhomogeneous otherwise.

Example 10.3.1

- In the homogeneous case, if we assume y is never 0, we obtain $\frac{y'}{y} = -g(x)$.
- Integrating, this gives $\ln y = -G(X) + C$ for some $C \in \mathbb{R}$ if $y > 0$ and $\ln(-y) = -G(x)\,\mathrm{d}x + C$ if $y < 0$, where G is some anti-derivative of g, i.e., $\int g(x)\,\mathrm{d}x = G(x) + C$.
- Hence $y = \pm e^C e^{G(x)}$ and so in either case, $y = Ke^{-G(x)} = Ke^{-\int g(x)\,\mathrm{d}x}$ for some $K \neq 0$.
  - As usual, we can now check directly that this is indeed a solution.
  - (And we notice that $K = 0$ also yields a solution.)

6 Feb 12, More linear equations, and linear algebra review

6.1 10.3A: Exponential integrating factors

- In order to show that the solution $y = Ke^{-G(x)}$ is the *general* solution (which we don't quite know since we had to assume $y \neq 0$), we use a trick we have used before, and divide the original equation by the known solution $y = e^{-G(x)}$.
  - We obtain $0 = e^{G(x)}y' + e^{G(X)}g(x) = \frac{\mathrm{d}}{\mathrm{d}x}(e^{G(x)}y)$.
  - Note that this is equivalent to the original equation, since we have multiplied by both sides by a non-vanishing quantity.
  - Thus, we see that for any solution y to this equation, $e^{G(x)y}$ must be constant, and so $y = Ke^{-G(x)}$ for some $K \in \mathbb{R}$, as desired.
- This is the trick of "exponential integrating factors" and it also allows us to solve linear equations in the *inhomogeneous case*.
- For the equation $y' + g(x)y = f(x)$ in normalized form, the exponential integrating factor is $M(x) = e^{\int g(x)\,\mathrm{d}x}$.
- We multiply the equation by M and obtain

$$e^{\int g(x)\,\mathrm{d}x}y' + g(x)e^{\int g(x)\,\mathrm{d}x} = f(x)e^{\int g(x)\,\mathrm{d}x}.$$

- Using the product rule, the left-hand side is $\frac{\mathrm{d}}{\mathrm{d}x}(ye^{\int g(x)\,\mathrm{d}x})$, and the equation becomes

$$\frac{\mathrm{d}}{\mathrm{d}x}(ye^{\int g(x)\,\mathrm{d}x}) = f(x)e^{\int g(x)\,\mathrm{d}x}.$$

- Now we can just integrate the right-hand side and solve for y.
  - Again, the original equation is equivalent to the equation after multiplying by M since M is nowhere zero.

Example 10.3.2

- Let's solve $y' = xy + x$.
- We rewrite it in normalized form $y' - xy = x$.
- Multiply by the integrating factor $y'e^{-x^2/2} - xye^{-x^2/2} = xe^{-x^2/2}$.
- Now integrate $ye^{-x^2/2} = -e^{-x^2/2} + C$.
- Thus $y = -1 + Ce^{-x^2/2}$.
- Since each step was an equivalence, this is the *general* solution to the equation (with domain $\mathbb{R}$).

6.2 10.3B: Applications

Example 10.3.5

- Newton's law of cooling says the surface temperature $u(t)$ of an objects changes at a rate proportional to the difference to the ambient temperature $f(t)$ (which we are assuming might also vary with time).
  - Thus $u' = k(f - u)$ for some $k > 0$.

  - k should be positive so that $u' < 0$ precisely when $f < u$.

- It is supposedly quite accurate in certain situations, but as with all empirical laws, it has its limitations.
  - Ideally, in addition to testing the law with experiments, one should develop a *mechanism* (for example, in terms of the jiggling of molecules) which explain why it should hold; this makes it easier to understand under which circumstances it should and shouldn't apply.

- Writing this as $u' + ku = kf(t)$, this is just a special case of a linear first-order equation in which the coefficient of u is constant.

- We obtain $u'e^{kt} + ue^{kt} = kf(x)e^{kt}$ and hence $ue^{kt} = \int kf(x)e^{kt}\,\mathrm{d}t + C$ and hence $u = ke^{-kt}\int f(t)e^{kt}\,\mathrm{d}t + Ce^{-kt}$.
  - We fix the lower-bound of the integral to $t = 0$, giving $C = u(0)$ and hence $u = ke^{-kt}\int_0^t f(s)e^{ks}\,\mathrm{d}s + u(0)e^{-kt}$.

- If the ambient temperature is constant, $f(t) = f_0$, we obtain $u = kf_0(1 - e^{-kt}) + u(0)e^{-kt} = kf_0 + (u(0) - kf_0)e^{-kt}$.
  - Thus $u(t) \xrightarrow{t\to\infty} f_0$, as we would expect.

6.3 Abstract vector spaces

- We review the concept of an abstract vector space.

- There are both real and complex vector spaces. In order to handle both at the same time, in what follows, we let K stand for one of $\mathbb{R}$ or $\mathbb{C}$.
  - (Actually, the notion of vector space makes sense when K is any *field*.)

- The motivating examples of vector spaces are the familiar ones K^n; the abstract definition results from isolating the most important features and properties from these. One should keep this example in mind when considering the general definition, but as we'll see, there are lots of other interesting examples as well.

- **Definition**: a vector space over K or K-vector space is a triple $(V, +, \times)$ where V is an arbitrary *set* (whose elements we call *vectors*), $+\colon V \times V \to V$ is a binary operation on V (called "*addition*") taking any two elements $\mathbf{u}, \mathbf{v} \in V$ to an element $\mathbf{u} + \mathbf{v} \in V$, and $\times\colon K \times V \to V$ is an operation (called "*scalar multiplication*") taking an element $r \in K$ and an element $\mathbf{v} \in V$ to an element $r \times \mathbf{v} \in V$ (which we also write as $r \cdot \mathbf{v}$ or just $r\mathbf{v} \in V$).
  - These are required to satisfy the following *axioms*:
  - Addition is associative and commutative, and has an identity element $\mathbf{0} \in V$ (i.e., $\mathbf{v} + \mathbf{0} = \mathbf{v}$ for all $\mathbf{v} \in V$). (*Exercise*: it follows that there is a *unique* such element, which we call the "*zero vector*" of V.)
  - Each $\mathbf{v} \in V$ has an additive inverse $(-\mathbf{v})$ (i.e., $\mathbf{v} + (-\mathbf{v}) = \mathbf{0}$). (*Exercise*: it follows that for each $\mathbf{v}$, there is a *unique* such element $-\mathbf{v}$.)
  - Scalar multiplication distributes over vector addition and scalar addition, i.e., $r \cdot (\mathbf{u} + \mathbf{v}) = r \cdot \mathbf{u} + r \cdot \mathbf{v}$ and $(r + s) \cdot \mathbf{v} = r \cdot \mathbf{v} + s \cdot \mathbf{v}$ for all $r, s \in K$ and $\mathbf{u}, \mathbf{v} \in V$.
  - $1 \cdot \mathbf{v} = \mathbf{v}$ for all $\mathbf{v} \in V$.
  - $(r \cdot s) \cdot \mathbf{v} = r \cdot (s \cdot \mathbf{v})$ for all $r, s \in K$ and $\mathbf{v} \in V$.

- As is typical, we will often *abuse notation* and simply write V in place of $(V, +, \times)$.
  - For example, we may speak of "the vector space $\mathbb{R}^3$", whereas the vector space is really the triple $(\mathbb{R}^3, +, \times)$.
- Examples:
  - The original and most important example of a K-vector space is the set K^n of n-tuples of elements of K. Addition and multiplication are given component-wise, i.e., $(u_1, \ldots, u_n) + (v_1, \ldots, v_n) = (u_1 + v_1, \ldots, u_n + v_n)$, and $r \cdot (v_1, \ldots, v_n) = (rv_1, \ldots, rv_n)$ for $r \in K$.
    - (One should check that these operations indeed satisfy the axioms!)
    - Of course, we usually picture $\mathbb{R}^2$ as the set of points in a plane with a fixed pair of coordinate axes and chosen units of length, and likewise $\mathbb{R}^3$ is pictured as 3-dimensional space (and $\mathbb{R}^1$ as a line).
    - (However, when considering them as vector spaces, we should really regard $\mathbb{R}^2$, for instance, as the set of *arrows* in the given plane, where we identify two arrows if one can be brought on top of the other by a translation; the reason is that it doesn't make sense to add or scale *points*, but we do know how to add *arrows* (in the usual head-to-tail manner) and scale them. The identification between arrows and points of course follows from placing the tail an arrow at the origin and passing to the point at its head.)
  - A related example is the set $K^{m \times n}$ of $m \times n$-matrices. Formally, this is the set of functions

$$\{(i, j) \mid 1 \leq i \leq m; 1 \leq j \leq n\} \to K.$$

    - As with tuples, given a matrix A, we write A_{ij} in place of $A(i, j)$, and we represent a matrix in the familiar way as a box of numbers:

$$A = \begin{bmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & & \vdots \\ A_{m1} & \cdots & A_{mn} \end{bmatrix}.$$

    - We also write $A = (A_{ij})_{1 \leq i \leq m, j \leq 1 \leq n}$ or just $A = (A_{ij})$.
    - Addition and scalar multiplication are again given component-wise: $(A + B)_{ij} = A_{ij} + B_{ij}$ and $(r \cdot A)_{ij} = r \cdot A_{ij}$ for $r \in K$.
    - Of special interest are the vector spaces $K^{n,1}$ of column vectors and $K^{1,n}$ of row vectors of length n. Of course, there are bijections $K^n \cong K^{n,1} \cong K^{1,n}$, and we may sometimes *abuse notation* and simply write K^n in place of $K^{n,1}$.
  - The above two examples generalize: given any set S, the set K^S of all functions $S \to K$ is a vector space with operations $(f + g)(s) = f(s) + g(s)$ and $(r \cdot f)(s) = r \cdot f(s)$.
    - In particular, the set $\mathbb{R}^{\mathbb{R}}$ of all functions $\mathbb{R} \to \mathbb{R}$ becomes a vector space in this way.

7 Feb 17, More linear algebra

7.1 Subspaces

- A linear subspace (or just subspace) of a vector space V is a subset $W \subset V$ which is closed under addition and scalar multiplication:
  - This means that $\mathbf{u}, \mathbf{v} \in W$ implies $\mathbf{u} + \mathbf{v} \in W$ and $r\mathbf{v} \in W$ for all $r \in K$.
  - In this case, W is again a vector space, with the same operations $+$ and $\times$.
- The most familiar examples are the subspaces of $\mathbb{R}^3$, which are the planes and lines passing through the origin, as well as $\mathbb{R}^3$ itself and the singleton set $\{\mathbf{0}\} \subset \mathbb{R}^3$.
- Other interesting examples of vector spaces arise as subspaces of $\mathbb{R}^{\mathbb{R}}$ (or more generally of $\mathbb{R}^I$ for any domain $I \subset \mathbb{R}$):
  - The set of *continuous* functions $\mathcal{C}^0(I)$ and the set of *smooth* functions $\mathcal{C}^\infty(I)$ are each subspaces of $\mathbb{R}^{\mathbb{R}}$ (as is $\mathcal{C}^k(I)$ for any k).
  - The set of *polynomial functions* $p(x) = \sum_{i=0}^n a_i x^i$ (with $n \in \mathbb{N}$ and $a_i \in \mathbb{R}$) is a subspace of $\mathbb{R}^{\mathbb{R}}$, and is usually denoted $\mathbb{R}[x]$.
  - There is a further subspace $\mathbb{R}_{\leq d}[x]$ consisting of polynomials of degree at most d.
  - Similarly, one can consider polynomials $\mathbb{C}[x]$ (and $\mathbb{C}_{\leq d}[x]$) over $\mathbb{C}$.
  - One should contemplate why these are all in fact subspaces!

7.2 Basis, dimension, etc.

- Given vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ in a vector space V, a linear combination of these vectors is any vector of the form $\sum_{i=1}^n a_i \mathbf{v}_i = a_1\mathbf{v}_1 + \cdots + a_n\mathbf{v}_n$ with $a_1, \ldots, a_n \in K$.
- The span $\mathrm{Span}(S)$ of a subset $S \subset V$ is the set of all linear combinations $\sum_{i=1}^n a_i \mathbf{v}_i$ with $\mathbf{v}_1, \ldots, \mathbf{v}_n \in S$.
- We say that S spans V or is a spanning set if $\mathrm{Span}(S) = V$, i.e., if every element of V is a linear combination of elements in S.
- We say that S is linearly independent if, whenever $\mathbf{v}_1, \ldots, \mathbf{v}_n \in S$ are distinct elements and $a_1\mathbf{v}_1 + \cdots + a_n\mathbf{v}_n = \mathbf{0}$, then $a_1 = \cdots = a_n = 0$
  - Equivalently, each element of V can be represented *in at most one way* as a linear combination of elements of S.
  - Equivalently, no element of S is a linear combination of the other elements.
- A basis for V is a subset $\mathcal{B}$ which is both spanning and linearly independent (or equivalently, such that every vector in V can be represented in *exactly one way* as a linear combination of elements of S).
  - K^n has the standard basis $\mathbf{e}_1, \ldots, \mathbf{e}_n$, where $\mathbf{e}_i$ is the vector $\mathbf{e}_i = (0, \ldots, 1, \ldots, 0)$ with a single 1 in the i-th place and 0 elsewhere.
  - It is often useful to regard a basis not as a *set* $\mathcal{B} \subset V$, but as a *tuple* $B \in V^n$; in other words, we remember the *order* of the elements in B, and in this case refer to B as an ordered basis.
  - (However, we may sometimes abuse terminology and simply say "basis" when we mean "ordered basis".)

- If V has a basis $\mathcal{B}$ which is *finite*, then it is called finite-dimensional; otherwise, it is called infinite-dimensional.
- **Theorem**: if V is finite-dimensional, then any two bases of V have the *same* number of elements, and this number is called the *dimension* $\dim(V)$ of V; moreover, any *linearly independent* set of vectors in V has at most $\dim V$ elements.
  - Roughly speaking, the proof proceeds by taking an arbitrary linearly independent set $S = \{\mathbf{u}_1, \ldots, \mathbf{v}_k\} \subset V$ and an arbitrary spanning set $T = \{\mathbf{v}_1, \ldots, \mathbf{v}_l\} \subset V$, and swapping out vectors in T for vectors in S one by one until a new spanning set is produced which includes all the vectors in S. This proves that $k \leq l$ for any such sets S and T (and in particular that $k = l$ if S and T are both bases).
- Examples of finite-dimensional vector spaces:

V	K^n	$K^{m\times n}$	$K_{\leq d}[x]$
$\dim(V)$	n	$m \cdot n$	$d+1$

  - Examples of infinite-dimensional vector spaces: $\mathbb{R}^{\mathbb{R}}, \mathcal{C}^0(\mathbb{R}), \mathcal{C}^\infty(\mathbb{R}), \mathbb{R}[x]$.
- **Lemma**: given any set $\mathbf{v}_1, \ldots, \mathbf{v}_k$ of linearly independent vectors in a vector space V, and a vector $\mathbf{v}_{k+1} \notin \mathrm{Span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$, the set $\{\mathbf{v}_1, \ldots, \mathbf{v}_{k+1}\}$ is still linearly independent.
- **Corollary** (of the theorem and the lemma): if V is finite-dimensional, then any subspace $W \subset V$ is finite-dimensional, and if $\dim W = \dim V$, then $W = V$.
  - Indeed, if W were infinite-dimensional, we would have more than $\dim V$ linearly independent vectors in V.
  - And if $\dim W = \dim V$ and $W \subsetneq V$, then by the Lemma, we could extend a basis of W to a larger linearly independent set of vectors, which would thus have more than $\dim V$ elements.
- Another **Corollary**: any linearly independent set of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k \in V$ in a finite-dimensional space can be extended to a basis $\mathbf{v}_1 \ldots, \mathbf{v}_n$ of V.
  - Indeed, as long as $k < n$, we must have $W = \mathrm{Span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\} \subsetneq V$, hence we can find some v_{k+1} in $V - W$, and the set $\{\mathbf{v}_1, \ldots, \mathbf{v}_{k+1}\}$ is still linearly independent; eventually, we will have $k = n$, and then $\mathbf{v}_1, \ldots, \mathbf{v}_n$ will have to constitute a basis.

8 Feb 19, Linear maps

8.1 Linear maps

- A function $f\colon V \to W$ between two vector spaces is linear if it preserves addition and scalar multiplication, i.e., $f(\mathbf{v}+\mathbf{w}) = f(\mathbf{v}) + f(\mathbf{w})$ and $f(r\mathbf{w}) = r \cdot f(\mathbf{w})$ for all $r \in K$.
  - It follows (by induction) that f preserves arbitrary linear combinations, i.e., that $f(a_1\mathbf{v}_1 + \cdots + a_n\mathbf{v}_n) = a_1(\mathbf{v}_1) + \cdots + a_n f(\mathbf{v}_n)$ for any $\mathbf{v}_1, \ldots, \mathbf{v}_n \in V$ and $a_1, \ldots, a_n \in K$.
  - A linear map $f\colon V \to V$ from a vector space to itself is sometimes called a linear operator.
- Given a basis $\mathbf{b}_1, \ldots, \mathbf{b}_n$ of V, any linear map $f\colon V \to W$ is determined by its values $f(\mathbf{b}_i)$ on the basis elements.
  - And conversely, given any vectors $\mathbf{w}_1, \ldots, \mathbf{w}_n$, there is a unique linear map $f\colon V \to W$ with $f(\mathbf{b}_i) = \mathbf{w}_i$ for all i; it is given by $f\left(\sum_{i=1}^n v_i\mathbf{b}_i\right) = \sum_{i=1}^n v_i\mathbf{w}_i$.
- *Exercise*: the linear maps $f\colon K^m \to K^n$ are exactly those of the form

$$f\begin{bmatrix} v_1 \\ \vdots \\ v_m \end{bmatrix} = \begin{bmatrix} a_{11}v_1 + \cdots + a_{1n}v_n \\ \vdots \\ a_{m1}v_1 + \cdots + a_{mn}v_n \end{bmatrix}$$

  for some elements $a_{ij} \in K$.
  - In other words, we have $f(\mathbf{v}) = A \cdot \mathbf{v}$, where $A = (a_{ij})$.
  - Here, we are using matrix multiplication, which we recall is the operation $K^{l\times m} \times K^{m\times n} \to K^{l\times n}$ given by $(A \cdot B)_{ij} = \sum_{k=1}^m A_{ik} \cdot B_{kj}$.
- For any two vector spaces V and W, we write $\mathcal{L}(V, W)$ for the set of all linear maps $V \to W$; this is a subspace of the vector space W^V.
  - In the case $V = W$, we simply write $\mathcal{L}(V)$ for the space $\mathcal{L}(V, V)$ of *linear operators* on V.
- Another important example of a linear map is the derivative operator $D = \frac{\mathrm{d}}{\mathrm{d}x}\colon \mathcal{C}^\infty(I) \to \mathcal{C}^\infty(I)$; this can also be considered as a linear map $\mathbb{R}_{\leq d}[x] \to \mathbb{R}_{\leq d-1}[x]$ or for any $d > 0$.
  - Similarly, we have seen the first-order *linear differential operators* $L\colon \mathcal{C}^\infty(I) \to \mathcal{C}^\infty(I)$ given by $L(y) = Dy + g(x)y = y' + g(x)y$ for some $g\colon I \to \mathbb{R}$.
  - There is also the integral map $\int_0^x\colon \mathcal{C}^0(\mathbb{R}) \to \mathcal{C}^0(\mathbb{R})$.
- The kernel or nullspace of a linear map $f\colon V \to W$ is the subspace $\ker(f) \subset V$ defined by $\ker(f) = \{\mathbf{v} \in V \mid f(\mathbf{v}) = \mathbf{0}\}$.
  - *Exercise*: a linear map is injective if and only if its nullspace is $\{\mathbf{0}\}$.
- The image $\operatorname{im}(f)$ of a linear map $f\colon V \to W$ (or in fact of any function between two sets) is the set $\operatorname{im}(f) = \{f(\mathbf{v}) \mid \mathbf{v} \in V\}$.
- **Theorem** (the *rank-nullity theorem* or *dim sum theorem*): if V and W are finite-dimensional and $f\colon V \to W$ is linear, then $\dim(V) = \dim\ker(f) + \dim\operatorname{im}(f)$.
  - In particular, if $\dim V = \dim W$, then f is injective if and only if it is surjective.
- An isomorphism between vector spaces V and W is a linear map $V \to W$ which is also a bijection.

  - In this case, the inverse map $W \to V$ is also linear (hence also an isomorphism).
  - If there exists an isomorphism between V and W, we say that V and W are isomorphic, and write $V \cong W$.

- Note that isomorphism is an *equivalence relation*: it is reflexive ($V \cong V$), symmetric ($V \cong W \Rightarrow W \cong V$), and transitive $U \cong V \wedge V \cong W \Rightarrow U \cong W$.

- **Theorem**: V is of finite-dimension n if and only if $V \cong K^n$.

  - Indeed, if V has a basis $\mathbf{b}_1, \dots, \mathbf{b}_n$, then each vector $\mathbf{v} \in V$ has *coordinates* $v_1, \dots, v_n$ with respect to this basis, determined by $\mathbf{v} = v_1\mathbf{b}_1 + \cdots v_n\mathbf{b}_n$; the map $V \to K^n$ taking $\mathbf{v}$ to $(v_1, \dots, v_n)$ is then the desired isomorphism.
  - An important consequence of this is that any linear map $f\colon V \to W$ between finite-dimensional vector spaces (say, with $\dim V = m$ and $\dim W = n$) can be represented by a $n \times m$ matrix, because $V \cong K^m$ and $W \cong K^n$.
    * The matrix representation of f will depend on a choice of isomorphisms $V \xrightarrow{\sim} K^m$ and $W \xrightarrow{\sim} K^n$, which is to say on a choice of ordered bases of V and W.
    * Given such ordered bases $\mathcal{B} = (\mathbf{v}_1, \dots, \mathbf{v}_m)$ and $\mathcal{C} = (\mathbf{w}_1, \dots, \mathbf{w}_n)$, the resulting matrix $A \in K^{n\times m}$ can be described directly by the formula $f(\mathbf{v}_i) = \sum_{j=1}^{n} A_{ji}\mathbf{w}_j$ (i.e., the i-th column of A consists of the coordinates of $f(\mathbf{v}_i)$ with respect to $\mathcal{C}$).
    * Another way to say this is: if we write $\mathbf{v}_\mathcal{B} \in K^{n\times 1}$ for the coordinate vector of $\mathbf{v} \in V$ with respect to the ordered basis $\mathcal{B}$, and similarly write $\mathbf{w}_\mathcal{C} \in K^{m\times 1}$ for the coordinates of $\mathbf{w} \in W$ with respect to $\mathcal{C}$, then we have $\big(f(\mathbf{v})\big)_\mathcal{C} = A \cdot \mathbf{v}_\mathcal{B} \in K^{m\times 1}$ for all $\mathbf{v} \in V$.

- Other examples of isomorphisms:

  - If $\dim V = m$ and $\dim W = n$, then $\mathcal{L}(V, W) \cong K^{n\times m}$ and in particular $\mathcal{L}(V) \cong K^{m\times m}$.
  - The existence of these isomorphisms follows from the above theorem just by comparing dimensions, but one can also establish the isomorphisms more directly by sending each linear map to the matrix representing it.
    * But, again, it is important to remember that this isomorphism will depend on choosing bases for V and W.
  - *Important exercise:* The above isomorphism takes *composition of linear maps* to *matrix multiplication*: given finite-dimensional spaces U, V, W with chosen ordered bases, and linear maps $f\colon U \to V$ and $g\colon V \to W$ represented by matrices $A \in K^{m\times l}$ and $B \in K^{n\times m}$, the composite map $g \circ f\colon U \to W$ is represented by the product $B \cdot A$.

9 Feb 24, 3.7: Inner products

9.1 3.7A: General properties of inner products

- We introduce the notion of an *inner product* on a general vector space, which is a scalar-valued binary function $V \times V \to \mathbb{R}$ satisfying certain axioms.
  - While the concept of *abstract vector space* allows us to reproduce many of the geometric features of the motivating example $\mathbb{R}^n$, it does not allow us to define the notions of *length* or *angle*, which are of course crucial to studying geometry in $\mathbb{R}^2$ and $\mathbb{R}^3$; it is inner products that allows us to reproduce these notions in an arbitrary abstract vector space.
  - Just as the very notion of vector space is meant to capture the main features of the archetypical case $\mathbb{R}^n$, so the notion of inner product is meant to capture the main features of the dot product $\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^n x_i y_i$ on $\mathbb{R}^n$; again, this motivating example should be kept in mind while considering the definition, but it's also important to remember that there are other interesting examples as well.
- If U, V, W are K-vector spaces, a function $f\colon U \times V \to W$ is bilinear if it is linear in each argument, when the other argument is held fixed, i.e., $f(a\mathbf{u} + \mathbf{u}', \mathbf{v}) = af(\mathbf{u}, \mathbf{v}) + f(\mathbf{u}', \mathbf{v})$ and $f(\mathbf{u}, a\mathbf{v} + \mathbf{v}') = af(\mathbf{u}, \mathbf{v}) + f(\mathbf{u}, \mathbf{v}')$ for all $\mathbf{u} \in U$, $\mathbf{v} \in V$, and $a \in \mathbb{R}$.
  - (Examples: the cross-product $\times\colon \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}^3$, matrix multiplication $K^{a,b} \times K^{b,c} \to K^{a,c}$, composition of linear maps $\mathcal{L}(U, V) \times \mathcal{L}(V, W) \to \mathcal{L}(U, W)$.)
- A bilinear form on a K-vector space V is a bilinear function $\beta\colon V \times V \to K$.
- An inner product on a real vector space V is a bilinear form $\beta\colon V \times V \to \mathbb{R}$ which is moreover
  - symmetric, meaning $\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle$ for all $\mathbf{u}, \mathbf{v} \in V$
  - positive-definite, meaning that $\langle \mathbf{u}, \mathbf{u} \rangle > 0$ for all $\mathbf{u} \neq \mathbf{0}$
- An inner product space is a pair (V, β), where V is a real vector space, and $\beta\colon V \times V \to \mathbb{R}$ is an inner product on V.
  - (Again, we often abuse notation, and just write say that "V is a inner product space").
  - (In particular, by "the inner product space $\mathbb{R}^n$", we mean $\mathbb{R}^n$ equipped with the standard inner product.)
- **Notation**: for a given inner product $\beta\colon V \times V \to \mathbb{R}$, we will usually prefer to write $\langle \mathbf{u}, \mathbf{v} \rangle$ in place of $\beta(\mathbf{u}, \mathbf{v})$.
  - Accordingly, we may forego the additional name β, and instead, inserting placeholders "$-$" for $\mathbf{u}$ and $\mathbf{v}$ in the expression "$\langle \mathbf{u}, \mathbf{v} \rangle$", just refer to the inner product as $\langle -, - \rangle\colon V \times V \to \mathbb{R}$.
- Remarks:
  - In light of the symmetry property, the bilinearity is equivalent to just being linear in one argument.
  - It follows from bilinearity that $\langle \mathbf{0}, \mathbf{0} \rangle = 0$, and hence (by positive-definiteness) that $\langle \mathbf{v}, \mathbf{v} \rangle \geq 0$ for all $\mathbf{v}$.
- Example: the bilinear forms on $\mathbb{R}^n$ are exactly the functions of the form $\langle \mathbf{u}, \mathbf{v} \rangle = \sum_{i,j=1}^n a_{ij} u_i v_j$ for some real numbers a_{ij}.

- In other words, $\langle u, v\rangle = \mathbf{u}^\top A\mathbf{v}$ where $A = (a_{ij})_{i,j=1}^n$, and where we are identifying $\mathbb{R}^{1\times 1}$ with $\mathbb{R}$.
- Recall here that for a matrix $B \in K^{l\times m}$, we write $B^\top \in K^{m\times l}$ for its transpose defined by $(B^\top)_{ij} = B_{ji}$.
- In particular, if we take A to be the identity matrix I_n (defined by $(\mathrm{I}_n)_{ij} = \delta_{ij}$), we obtain the dot product $\mathbf{u}\cdot\mathbf{v}$, also known as the standard inner product on $\mathbb{R}^n$.
  * The symbol δ_{ij}, called the Kronecker delta symbol, is defined to be 1 if $i = j$ and 0 otherwise
  * Often we will just write I instead of I_n when n is clear from context.

- It is easy to see that such a bilinear form $\mathbf{u}^\top A\mathbf{v}$ is symmetric precisely when the *matrix A is* symmetric, meaning $A^\top = A$.
- Example 3.7.1
  - It is more difficult to determine whether the bilinear form arising from a matrix A is *positive-definite*.
  - But here is an example: if A is a diagonal matrix (meaning $A_{ij} = 0$ for $i \neq j$) with *positive* diagonal entries, then $\mathbf{u}^\top A\mathbf{v}$ is positive definite (and hence an inner product, since A is clearly symmetric).
  - In this case $\langle \mathbf{u}, \mathbf{v}\rangle$ is given by the simple formula $a_{11}u_1v_1 + \cdots + a_{nn}u_nv_n$.
  - For a completely specific example, $\langle \mathbf{u}, \mathbf{v}\rangle = u_1u_2 + 2v_1v_2$ is an inner product on $\mathbb{R}^2$.

Complex inner products

- There is also a notion of inner products over the complex numbers, the standard example being the function $\mathbb{C}^n \times \mathbb{C}^n \to \mathbb{C}$ given by $(\mathbf{w}, \mathbf{z}) \mapsto \overline{\mathbf{w}}\cdot\mathbf{z}$.
  - Here, we recall that for a complex number $z = a + bi$, its complex conjugate is defined by $\bar{z} = a - bi$ (reflection across the real axis).
  - This has the fundamental property $z\cdot\bar{z} = |z|^2$, where we recall that the absolute value (or *modulus*) of a complex number is defined by $|z| = \sqrt{a^2+b^2}$ (distance from the origin in the complex plane).
  - Another fundamental property of the complex conjugate is that it is compatible with addition and multiplication: $\overline{z+w} = \bar{z} + \bar{w}$ and $\overline{z\cdot w} = \bar{z}\cdot\bar{w}$.
  - For a vector $\mathbf{w} \in \mathbb{C}^n$, the complex conjugate is defined component-wise: $\overline{\mathbf{w}} = (\overline{w}_1, \ldots, \overline{w}_n)$.
  - We recall the important formulas $\operatorname{Re} z = \frac{1}{2}(z + \bar{z})$ and $\operatorname{Im} z = \frac{1}{2}(z - \bar{z})$, where Re and Im are the real and imaginary parts, defined by $\operatorname{Re}(a+bi) = a$ and $\operatorname{Im}(a+bi) = b$.

- In general, on a complex vector space V, we define a function $\langle -, -\rangle \colon V \times V \to \mathbb{C}$ to be:
  - sesquilinear if it is linear in its second argument and *antilinear* in its first argument, meaning $\langle a\mathbf{u} + \mathbf{v}, \mathbf{w}\rangle = \bar{a}\langle \mathbf{u}, \mathbf{w}\rangle + \langle \mathbf{v}, \mathbf{w}\rangle$ and $\langle \mathbf{w}, a\mathbf{u} + \mathbf{v}\rangle = a\langle \mathbf{w}, \mathbf{u}\rangle + \langle \mathbf{w}, \mathbf{v}\rangle$ for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$ and $a \in \mathbb{C}$.
  - conjugate symmetric if $\langle \mathbf{u}, \mathbf{v}\rangle = \overline{\langle \mathbf{v}, \mathbf{u}\rangle}$ for all $\mathbf{u}, \mathbf{v} \in V$.
  - an inner product if it is sesquilinear, conjugate symmetric, and positive definite (which as before means $\langle \mathbf{v}, \mathbf{v}\rangle > 0$ for all $\mathbf{v} \neq \mathbf{0}$).
  - Again, a complex inner product space is a complex vector space equipped with an inner product.

- Note that if V is a *real* vector space, and $\langle -, - \rangle \colon V \times V \to \mathbb{R}$ is a *real*-valued function, then the definitions of sesquilinear and conjugate symmetric reduce to those of "bilinear" and "symmetric" (since $\bar{a} = a$ for all $a \in \mathbb{R}$).
  - Thus, for conciseness, we can define an inner product on a real *or* complex vector V all at once by saying it is a sesquilinear, conjugate symmetric, positive definite function $V \times V \to K$.

Norms

- Now let V be a real or complex inner product space.
- We define the norm or length of a vector $\|\mathbf{v}\|$ to be the non-negative real number $\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}$.
  - Note that $\|\mathbf{v}\| = 0 \iff \mathbf{v} = \mathbf{0}$ by positive-definiteness of $\langle -, - \rangle$.
  - Some people just write $|\mathbf{v}|$ for the norm instead of $\|\mathbf{v}\|$.
- The norm has the following properties:
  - ("Positivity") $\|\mathbf{v}\| > 0$ for all $\mathbf{v} \neq \mathbf{0}$
  - ("Linearity") $\|a\mathbf{v}\| = |a| \cdot \|\mathbf{v}\|$ for all $a \in K$ and $\mathbf{v} \in V$.
  - ("Triangle inequality") $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$
  - (In general, one calls a function $V \to \mathbb{R}$ having these properties a norm on V; there are examples of norms that do not arise from inner products, an example being $\|-\|_3 \colon \mathbb{R}^n \to \mathbb{R}$ given by $\|\mathbf{v}\|_3 = (\sum_{i=1}^n v_i^3)^{1/3}$.)
- We have the important formula (in the real case): $\|u + v\|^2 = \|u\| + \|v\| + 2\langle u, v \rangle$.
  - In the complex case, the formula is $\|u + v\|^2 = \|u\| + \|v\| + 2\,\mathrm{Re}\langle u, v \rangle$
  - These formulas show that the inner product can be recovered from the norm, namely (in the real case) as $\langle u, v \rangle = \frac{1}{2}(\|u + v\| - \|u\| - \|v\|)$.
- **Theorem (the Cauchy-Schwartz inequality)**: $|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}\| \cdot \|\mathbf{v}\|$ for all $\mathbf{u}, \mathbf{v} \in V$, where equality holds if and only if $\mathbf{u}$ and $\mathbf{v}$ are colinear (i.e., linearly dependent).
- **Proof**:
  - We give the proof in the real case; the complex case is very similar but a little bit more complicated.
  - The case in which $\mathbf{u} = \mathbf{0}$ or $\mathbf{v} = \mathbf{0}$ is obvious, since then both sides are equal to 0 (and $\mathbf{u}, \mathbf{v}$ are colinear in this case).
  - Next, assume $\mathbf{u}, \mathbf{v} \neq 0$. Using bilinearity of the inner product, and the linearity of the norm, we may divide both sides by $\|\mathbf{u}\| \cdot \|\mathbf{v}\|$ to obtain the equivalent inequality $|\langle \mathbf{a}, \mathbf{b} \rangle| \leq \|\mathbf{a}\| \cdot \|\mathbf{b}\|$, where $\mathbf{a} = \mathbf{u}/\|\mathbf{u}\|$ and $\mathbf{b} = \mathbf{v}/\|\mathbf{v}\|$.
  - We observe that $\|\mathbf{a}\| = \|\mathbf{b}\| = 1$; moreover, $\mathbf{a}, \mathbf{b}$ are colinear if and only if $\mathbf{u}, \mathbf{v}$ are. In other words, we have reduced to the case of proving the inequality for unit vectors.
  - In this case, we have $0 \leq \|\mathbf{a} \pm \mathbf{b}\|^2 = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 \pm 2\langle \mathbf{a}, \mathbf{b} \rangle = 2 \pm 2\langle \mathbf{a}, \mathbf{b} \rangle$, and hence $\pm\langle \mathbf{a}, \mathbf{b} \rangle \leq 1 = \|\mathbf{a}\| \cdot \|\mathbf{b}\|$, as desired.
  - Note that the inequality above is an equality if and only if $\mathbf{a} \pm \mathbf{b} = 0$.
- Example 3.7.1 again

- Note that for the inner product $\langle \mathbf{u}, \mathbf{v} \rangle = u_1 v_1 + 2u_2 v_2$ on $\mathbb{R}^2$, the resulting geometry in the plane is "distorted". For example, $(1, 0)$ is still a unit vector, but $(0, 1)$ now has length $\sqrt{0^2 + 2 \cdot 1^2} = \sqrt{2}$.
- The "unit circle" (the set of all vectors of length 1) is now an ellipse.

Angles

- We continue to fix a real or complex inner product space V.
- For $\mathbf{u}, \mathbf{v} \neq \mathbf{0}$, the Cauchy-Schwartz inequality can be rephrased as $-1 \leq \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{u}\| \|\mathbf{v}\|} \leq 1$ for any $\mathbf{u}, \mathbf{v} \in V$.
- It follows that there is a unique $\theta \in [0, \pi]$ with $\cos \theta = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{u}\| \|\mathbf{v}\|}$.
- We define the angle $\angle(\mathbf{u}, \mathbf{v})$ between $\mathbf{u}$ and $\mathbf{v}$ to be this number θ.
  - This recovers the usual notion of angle in $\mathbb{R}^2$ and $\mathbb{R}^3$ because of the well-known formula $\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\| \|\mathbf{v}\| \cos \theta$.
  - $\mathbf{u}$ and $\mathbf{v}$ are called orthogonal, written $\mathbf{u} \perp \mathbf{v}$, if $\theta = \pi/2$, or in other words if $\mathbf{u} \cdot \mathbf{v} = 0$.
  - This definition still makes sense when $\mathbf{u} = \mathbf{0}$ or $\mathbf{v} = \mathbf{0}$.
- We have the following version of the **Pythagorean theorem**:
  - If $\mathbf{u} \perp \mathbf{v}$, then $\|\mathbf{u} + \mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2$.
- Example 3.7.1 again
  - Note again that the non-standard inner product $\langle \mathbf{u}, \mathbf{v} \rangle = u_1 v_1 + 2u_2 v_2$ on $\mathbb{R}^2$ "distorts" the notion of angle.
  - For example, $(1, 0)$ and $(0, 1)$ are still orthogonal, but $(1, 1)$ and $(1, -1)$ are no longer orthogonal, since $\langle (1, 1), (1, -1) \rangle = 1 - 2 \neq 0$.

Example 3.7.2

- On the vector space $\mathcal{C}^0([0, 2\pi])$ of continuous real-valued functions on the interval $[0, 2\pi]$, we can define an inner product by the formula

$$\langle f, g \rangle = \int_0^{2\pi} f(x) g(x) \, \mathrm{d}x.$$

  - One should check that this is indeed an inner product.
- This inner product is of central importance in *Fourier analysis*, the mathematical subject underpinning of signal processing.
  - In this subject, it is also important to consider continuous complex-valued functions; in this case, the inner product can be defined by $\langle f, g \rangle = \int_0^{2\pi} \overline{f(x)} g(x) \, \mathrm{d}x$.

10 Feb 26, 3.7B: Orthogonal bases; dual spaces

10.1 3.7B: Orthogonal bases

- A set $S \subset V$ of vectors in a real or complex inner product space V is *orthonormal* if $\langle \mathbf{u}, \mathbf{u} \rangle = 1$ and $\langle \mathbf{u}, \mathbf{v} \rangle = 0$ for all $\mathbf{u}, \mathbf{v} \in V$ with $\mathbf{u} \neq \mathbf{v}$; i.e., all the vectors are unit vectors, and they are mutually orthogonal.
  - Any orthonormal set is linearly independent; indeed, if $a_1\mathbf{v}_1 + \cdots + a_n\mathbf{v}_n = \mathbf{0}$ where $\langle \mathbf{v}_i, \mathbf{v}_j \rangle = \delta_{ij}$ for all i, j, then using linearity of $\langle -, - \rangle$ and orthonormality, $0 = \langle \mathbf{v}_i, \mathbf{0} \rangle = \langle \mathbf{v}_i, a_1\mathbf{v}_1 + \cdots + a_n\mathbf{v}_n \rangle = a_i$ for all i.
- An orthonormal basis is a basis which is also an orthonormal set.
- The standard example of an orthonormal basis is the standard basis of $\mathbb{R}^n$ or $\mathbb{C}^n$ (equipped with the standard inner product).
- In general, a basis $\mathbf{b}_1, \ldots, \mathbf{b}_n$ of $\mathbb{R}^n$ is orthonormal if and only if the matrix B with columns $\mathbf{b}_1, \ldots, \mathbf{b}_n$ is an orthogonal matrix, meaning that $B^\top \cdot B = \mathrm{I}$.

We have the following nice formula for the representation of any vector with respect to an orthonormal basis.

Theorem 7.4

- If $\mathbf{b}_1, \ldots, \mathbf{b}_n$ is an orthonormal basis for an inner product space V, then $\mathbf{v} = \langle \mathbf{b}_1, \mathbf{v} \rangle \mathbf{b}_1 + \cdots + \langle \mathbf{b}_n, \mathbf{v} \rangle \mathbf{b}_n$ for any $\mathbf{v} \in V$.
- Proof:
  - Since $\mathbf{b}_1, \ldots, \mathbf{b}_n$ is a basis, we know that $\mathbf{v} = v_1\mathbf{b}_1 + \cdots + v_n\mathbf{b}_n$ for some uniquely determined $v_1, \ldots, v_n \in K$.
  - Using linearity and orthonormality (as in the above proof that orthonormal sets are linearly independent), we conclude that $\langle \mathbf{b}_i, \mathbf{v} \rangle = v_i$ for all i.

Similarly, we have a nice formula for the inner product of any two vectors represented in terms of an orthonormal basis.

Theorem 7.6

- If $\mathbf{b}_1, \ldots, \mathbf{b}_n \in V$ is an orthonormal basis for an inner product space V, then for any two vectors $\mathbf{x} = x_1\mathbf{b}_1 + \cdots + x_n\mathbf{b}_n$ and $\mathbf{y} = y_1\mathbf{b}_1 + \cdots + y_n\mathbf{b}_n$, their inner product is given by $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^n \bar{x}_i y_i$.
  - This follows immediately from (sesqui-)linearity and orthonormality.

Projection

- The above Theorem 7.4 can be understood geometrically as "any vector is the sum of its projections onto the vectors of any orthonormal basis"; to understand this, we need the notion of projection.
- If $\mathbf{u} \in V$ is a vector in an inner product space V, the (orthogonal) projection onto $\mathbf{u}$ is the linear map $\Pi_{\mathbf{u}} \colon V \to \operatorname{Span} \mathbf{u}$ given by $\Pi_{\mathbf{u}}\mathbf{v} = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle}\mathbf{u}$.
  - When $\mathbf{u}$ is a unit vector, this simplifies to $\Pi_{\mathbf{u}}\mathbf{v} = \langle \mathbf{u}, \mathbf{v} \rangle \mathbf{u}$.

- In this case, the *length* of the projection of $\mathbf{v}$ onto $\mathbf{u}$ is simply given by $|\langle \mathbf{u}, \mathbf{v} \rangle|$.
- To understand where this definition comes from, one should consider the case of $\mathbb{R}^2$ or $\mathbb{R}^3$, and draw a triangle and do some trigonometry. (The definition is also justified by Theorem 7.5 below.)

- More generally, let $\mathbf{b}_1, \ldots, \mathbf{b}_m \in V$ be any orthonormal set of vectors, and let $W = \mathrm{Span}(\mathbf{b}_1, \ldots, \mathbf{b}_m)$ (so that $\mathbf{b}_1, \ldots, \mathbf{b}_m$ is an orthonormal basis of W).
  - We define the (orthogonal) projection onto W to be the linear map $\Pi_W \colon V \to W$ given by $\Pi_W \mathbf{v} = \langle \mathbf{b}_1, \mathbf{v} \rangle \mathbf{b}_1 + \cdots \langle \mathbf{b}_m, \mathbf{v} \rangle \mathbf{b}_m$ (this should be compare to Theorem 7.4 above).

Theorem 7.5

- With $V, W, \mathbf{b}_1, \ldots, \mathbf{b}_m$ as above, for any $\mathbf{v} \in V$, the orthogonal projection $\Pi_W \mathbf{v} \in W$ is the unique closest vector to $\mathbf{v}$ in W, i.e., we have $\|\mathbf{v} - \Pi_W \mathbf{v}\| \leq \|\mathbf{v} - \mathbf{w}\|$ for any $\mathbf{w} \in W$, and if this is an equality, then $\mathbf{w} = \Pi_w \mathbf{v}$.

- In particular, this proves that the orthogonal projection map Π_W depends only on the finite-dimensional subspace W and not on the given choice of orthonormal basis.

- The proof rests on the following observation, which is of independent interest: that $\mathbf{v} - \Pi_W \mathbf{v} \perp$ W \) ("$\mathbf{v} - \Pi_W \mathbf{v}$ is orthogonal to W"), by which we mean $(\mathbf{v} - \Pi_W \mathbf{v}) \perp \mathbf{w}$ for every $\mathbf{w} \in W$.
  - This follows by linearity as soon as we show that $(\mathbf{v} - \Pi_W \mathbf{v}) \perp \mathbf{b}_i$ for every $\mathbf{b}_i$, and this follows immediately from sesquilinearity and orthonormality.

- Proof of the theorem:
  - Using the Pythagorean theorem and $(\mathbf{v} - \Pi_W \mathbf{v}) \perp W$, we have, for any $\mathbf{w} \in W$, that

$$\|\mathbf{v} - \mathbf{w}\|^2 = \|(\mathbf{v} - \Pi_W \mathbf{v}) + (\Pi_W \mathbf{v} - \mathbf{w})\|^2 = \|\mathbf{v} - \Pi_W \mathbf{v}\|^2 + \|\Pi_W \mathbf{v} - \mathbf{w}\|^2 \geq \|\mathbf{v} - \Pi_W \mathbf{v}\|^2.$$

  - Moreover, the above inequality is an equality if and only if $\|\Pi_W \mathbf{v} - \mathbf{w}\| = 0$, i.e., if and only if $\Pi_W \mathbf{v} = \mathbf{w}$; this proves uniqueness.

The Gram-Schmidt process

- What we showed above is that there exists an orthogonal projection operator Π_W onto any finite-dimensional subspace W *assuming W has an orthonormal basis.*

- We now recall the Gram-Schmidt process, which shows that **any finite-dimensional inner product space *does* have an orthonormal basis**.
  - More precisely, for any finite set of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_m \in V$ is an inner product space, the Gram-Schmidt process gives explicit formulas for a set of orthonormal vectors $\mathbf{b}_1, \ldots, \mathbf{b}_k$ having the same span.

- The process is to first produce a set of *orthogonal* vectors $\mathbf{w}_i$, by subtracting from each $\mathbf{v}_i$ the orthogonal projection onto the span of the previous vectors.
  - That is, we define $\mathbf{w}_1 = \mathbf{v}_1$, and then inductively define

$$\mathbf{w}_{i+1} = \mathbf{v}_i - \Pi_{\mathrm{Span}\{\mathbf{w}_1, \ldots, \mathbf{w}_i\}} \mathbf{v}_i = \mathbf{v}_i - \frac{\langle \mathbf{w}_1, \mathbf{v} \rangle}{\|\mathbf{w}_1\|^2} \mathbf{w}_1 - \cdots - \frac{\langle \mathbf{w}_i, \mathbf{v} \rangle}{\|\mathbf{w}_i\|^2} \mathbf{w}_i.$$

- If along the way, any $\mathbf{v}_i$ was dependent on the previous $\mathbf{v}_j$'s, then we get $\mathbf{w}_i = 0$, and in this case, we discard it (that is, we replace our original sequence $\mathbf{v}_1, \ldots, \mathbf{v}_m$ with the sequence $\mathbf{v}_1, \ldots, \mathbf{v}_{i-1}, \mathbf{v}_{i+1}, \ldots, \mathbf{v}_m$).
- We then define $\mathbf{b}_i = \frac{\mathbf{w}_i}{\|\mathbf{w}_i\|}$ to obtain an orthonormal set.

- (One can normalize the $\mathbf{w}_i$ to obtain the $\mathbf{b}_i$ "along the way" while performing the algorithm, instead of doing it at the end, but doing so usually makes the computations more complicated.)

10.2 3.7C: Orthogonal transformations

- A linear map $f\colon V \to W$ between inner product spaces is orthogonal if it preserves the inner product, i.e., $\langle f\mathbf{u}, f\mathbf{v}\rangle = \langle \mathbf{u}, \mathbf{v}\rangle$ for all $\mathbf{u}, \mathbf{v} \in V$.
  - Equivalently, $\|f\mathbf{v}\| = \|\mathbf{v}\|$ for all $\mathbf{v} \in V$ (this follows from the formula which recovers the inner product from the norm).
  - Equivalently, $\|f\mathbf{u} - f\mathbf{v}\| = \|\mathbf{u} - \mathbf{v}\|$ for all $\mathbf{u}, \mathbf{v} \in V$, i.e., f preserves the distances between points.

- If $\mathbf{b}_1, \ldots, \mathbf{b}_n$ is an orthonormal basis of V, then f is orthogonal if and only if $f(\mathbf{b}_1), \ldots, f(\mathbf{b}_n)$ is still an orthonormal set.
  - The $\Rightarrow$ direction is immediate.
  - The other direction follows because if $\mathbf{u} = \sum_i u_i\mathbf{b}_i$ and $\mathbf{v} = \sum_i v_i\mathbf{b}_i \in V$, then $\langle \mathbf{u}, \mathbf{v}\rangle = \sum_i \bar{u}_i v_i = \sum_{i,j} \bar{u}_i v_j \langle \mathbf{b}_i, \mathbf{b}_j\rangle$, and $\langle f\mathbf{u}, f\mathbf{v}\rangle = \sum_{i,j} \bar{u}_i v_j \langle f(\mathbf{b}_i), f(\mathbf{b}_j)\rangle$ using the linearity of f and sesquilinearity of the inner product.

- If V and W are real vector spaces, and we fix orthogonal bases of V *and* W, then $f\colon V \to W$ is orthogonal if and only if the matrix $A \in \mathbb{R}^{m\times n}$ representing f with respect to these bases satisfies $A^\top \cdot A = \mathrm{I}$.

- One can show that the orthogonal transformations $\mathbb{R}^2 \to \mathbb{R}^2$ and $\mathbb{R}^3 \to \mathbb{R}^3$ (equipped with the standard inner product) are exactly the usual rotations and reflections.

- An isomorphism of inner product spaces (or linear isometry) is a bijective orthogonal linear map.
  - We say that two inner product spaces $(V, \langle -, -\rangle_V)$ and $(W, \langle -, -\rangle_W)$ are isomorphic, denoted $(V, \langle -, -\rangle_V) \cong (W, \langle -, -\rangle_W)$ if there exists an isomorphism of inner product spaces $V \to W$.

- **Theorem**: any n-dimensional inner product space V over K is isomorphic to K^n (with the standard inner product).
  - Indeed, we have already seen that any choice $\mathbf{b}_1, \ldots, \mathbf{b}_n$ of basis for V determines a linear isomorphism $f\colon V \to K^n$ with $f(\mathbf{b}_i) = \mathbf{e}_i$, the i-th standard basis vector.
  - If this is moreover an *orthonormal basis*, then f is an orthogonal map, and hence an isomorphism of inner product spaces.

- Example 3.7.1 again
  - The theorem gives a new perspective on the "distorted" inner product $\langle \mathbf{u}, \mathbf{v}\rangle = u_1v_1 + 2u_2v_2$: we have an isomorphism $f\colon \mathbb{R}^2 \to \mathbb{R}^2$ taking the distorted inner product to the standard inner product, namely $f(v_1, v_2) = (v_1, \sqrt{2} \cdot v_2)$.
  - Hence, we can understand the distorted inner product $\langle \mathbf{u}, \mathbf{v}\rangle$ as being described by the procedure "first apply f, then take the ordinary inner product".

10.3 Dual spaces

- The dual space V^* of a vector space V over K is the vector space $\mathcal{L}(V, K)$ of all linear maps $V \to K$; elements of V^* are sometimes referred to as *covectors*, *dual vectors*, *functionals*, or *linear forms*.

- Each linear map $f\colon V \to W$ between vector spaces induces a dual map $f^*\colon W^* \to V^*$ by composition: for $\varphi \in W^*$, we set $f^*(\varphi) = \varphi \circ f \in V^*$.

  - The dual map f^* is linear.
  - Moreover, the operation $f \mapsto f^*$ determines a linear map $\mathcal{L}(V, W) \to \mathcal{L}(W^*, V^*)$.
  - Given linear maps $U \xrightarrow{f} V \xrightarrow{g} W$, we have $(g \circ f)^* = f^* \circ g^*\colon W^* \to U^*$.

- Any basis $\mathbf{b}_1, \ldots, \mathbf{b}_n$ of V gives rise to a dual basis $\varphi_1, \ldots, \varphi_n$ of V^*, determined by the conditions $\varphi_i(\mathbf{b}_j) = \delta_{ij}$ for all i, j.

  - That these conditions determine functionals φ_i follows from the fact that a linear map is completely specified by giving its values on any given basis.
  - To see that the φ_i form a basis: linear independence follows from the fact that if $\varphi = \sum_{i=1}^n a_i\varphi_i$, then $a_i = \varphi(\mathbf{b}_i)$ for all i; and the spanning property follows from the fact that for any $\varphi \in V^*$, we have $\varphi = \sum_{i=1}^n \varphi(\mathbf{b}_i)\varphi_i$.
    - Both these facts are established by evaluating both sides of the equation on all of the basis vectors $\mathbf{b}_i$.

- Dual maps are related to transposes of matrices in the following way:

  - Let V and W be vector spaces over K of dimensions $m = \dim V$ and $n = \dim W$, and fix bases $\mathcal{B}_V$ of V and $\mathcal{B}_W$ of W, and hence dual bases $\mathcal{B}_V^*$ and $\mathcal{B}_W^*$ of V^* and W^*. Given a linear map $f\colon V \to W$ represented by a matrix $A \in K^{n\times m}$ with respect to $\mathcal{B}_V$ and $\mathcal{B}_W$, the dual map $f^*\colon W^* \to V^*$ is represented by the matrix $A^\top$, with respect to the dual bases of $\mathcal{B}_W^*$ and $\mathcal{B}_V^*$.
  - To prove this, write $\mathcal{B}_V = (\mathbf{v}_1, \ldots, \mathbf{v}_m)$ and $\mathbf{w}_1, \ldots, \mathbf{w}_n$, and $\mathcal{B}_V^* = (\varphi_1, \ldots, \varphi_m)$ and $\mathcal{B}_W^* = (\psi_1, \ldots, \psi_n)$.
    - We then have by definition that $f(\mathbf{v}_i) = \sum_{j=1}^n A_{ji}\mathbf{w}_j$, and hence that $\psi_j(f(\mathbf{v}_i)) = A_{ji}$.
    - Similarly, if $B \in K^{m\times n}$ is the matrix representing f^* with respect to $\mathcal{B}_W^*$ and $\mathcal{B}_V^*$, then $f^*(\psi_j) = \sum_{i=1}^m B_{ij}\varphi_i$, and hence $(f^*(\psi_j))(\mathbf{v}_i) = B_{ij}$.
    - Hence
      $$B_{ij} = (f^*(\psi_j))(\mathbf{v}_i) = \psi_j(f(\mathbf{v}_i)) = A_{ji},$$
      so $B = A^\top$, as desired.

- **Lemma**: for any non-zero vector $\mathbf{v} \in V$ in a finite-dimensional vector space, there is some covector $\varphi \in V^*$ with $\varphi(\mathbf{v}) \neq 0$.

  - Indeed, we may extend $\{\mathbf{v}\}$ to a basis, and take φ to be the dual basis vector corresponding to $\mathbf{v}$, so that $\varphi(\mathbf{v}) = 1$.
  - (In fact, this Lemma also holds for infinite-dimensional spaces.)

- There is a natural linear map $\alpha\colon V \to V^{**}$ from a vector space to the dual of its dual space, given by $\alpha(v) = \mathrm{ev}_{\mathbf{v}}$, where $\mathrm{ev}_{\mathbf{v}}\colon V^* \to K$ is defined by $\mathrm{ev}_{\mathbf{v}}(\varphi) = \varphi(v)$.

  - By the Lemma, α is always injective, since if $\mathrm{ev}_{\mathbf{v}} = 0$, then $\varphi(\mathbf{v}) = \mathrm{ev}_{\mathbf{v}}(\varphi) = 0$ for every $\varphi \in V^*$.

  - If V is finite-dimensional, it follows that α is an isomorphism, since V and V^{**} have the same dimension.
  - That is, **the map α provides a canonical isomorphism between any finite-dimensional vector space and its double dual space**.

- Note that if V is finite-dimensional, we not only have $V \cong V^{**}$, but also $V \cong V^*$, simply for dimension reasons: $V \cong K^{\dim V} \cong V^*$.
  - However, unlike with V^{**}, there is no *canonical* such isomorphism; one has to *choose* bases of V and V^* in order to produce an isomorphism $V \to V^*$.
  - (Actually, it suffices to just choose a basis of V, since this induces a dual basis of V^*, and hence an isomorphism $V \xrightarrow{\sim} V^*$ taking each basis vector in V to the corresponding dual basis vector.)

- There is another important source of isomorphisms $V \xrightarrow{\sim} V^*$: on any real inner product space V there is a canonical (i.e., only depending on the inner product, but on no further choices) isomorphism $\rho\colon V \xrightarrow{\sim} V^*$, which takes a vector $\mathbf{v} \in V$ to the functional $\rho(\mathbf{v})\colon V \to \mathbb{R}$ defined by $\rho(\mathbf{v})(\mathbf{w}) = \langle \mathbf{v}, \mathbf{w} \rangle$.
  - ρ is injective by positive-definiteness: if $\rho(\mathbf{v}) = 0$, then $\langle \mathbf{v}, \mathbf{v} \rangle = \rho(\mathbf{v})(\mathbf{v}) = 0$ and hence $\mathbf{v} = 0$.
  - It then follows that ρ is an isomorphism since $\dim V = \dim V^*$.

- If $V = \mathbb{R}^{n,1}$ is the space of column vectors, then V^*, the space of linear maps $\mathbb{R}^{n,1} \to \mathbb{R} \cong \mathbb{R}^{1,1}$, is naturally identified with the space $\mathbb{R}^{1,n}$ of $(1 \times n)$-matrices, i.e., row vectors.
  - Under these identifications, the isomorphism $V \xrightarrow{\sim} V^*$ induced by the standard inner product on $\mathbb{R}^{n,1}$ is then simply the transposition map $\mathbb{R}^{n,1} \xrightarrow{\sim} \mathbb{R}^{1,n}$.
  - This happens to be the same as the isomorphism induced by taking the standard basis of $\mathbb{R}^{n,1}$ to its dual basis in $\mathbb{R}^{1,n}$.

11 Mar 3, 11.1: differential operators

11.1 11: Second-order equations

- We'll now start to study second order equations $y'' = F(x, y, y')$.
- Mostly we'll look at *linear* equations $y'' + f(x)y' + g(x)y = h(x)$, and more specifically, linear equations with *constant coefficients* $y'' + ay' + by = h(x)$ with $a, b \in \mathbb{R}$.
  - Some examples:

$$y'' - 2y = x$$
$$y'' - 3y = e^x$$
$$y'' - 3y' + 2y = f(x)$$

- Later, we'll also look at some *nonlinear* equations such as $y'' + (y')^2 = 0$.

11.2 11.1: Differential operators

- To solve a first-order linear equation $y' + g(x)y = f(x)$, we used the exponential multiplier $e^{\int g(x)\,\mathrm{d}x}$.
  - If the equation has constant coefficients $y' + ry = f(x)$ (as in Newton's law of cooling), this simply becomes e^{rx}.
  - We recall that if y is any solution, we have $e^{rx}(y' + ry) = e^{rx}f(x)$ and hence $\frac{\mathrm{d}}{\mathrm{d}x}(e^{rx}y) = e^{rx}f(x)$ and $y = Ce^{-rx} + e^{-rx}\int e^{rx}f(x)\,\mathrm{d}x$ for some $C \in \mathbb{R}$.
  - In what follows, we will be making further use of such exponential multipliers.
- Before proceeding, we introduce some notation
- For a given interval I, recall that we have the vector space $\mathcal{C}^\infty(I)$ of smooth functions on I.
  - We denote by $D\colon \mathcal{C}^\infty(I) \to \mathcal{C}^\infty(I)$ the *differentiation operator* given by $Dy = y'$.
  - We recall that D can alternatively be considered as having different domains: for example, it can regarded as an operator $D\colon \mathbb{R}[x] \to \mathbb{R}[x]$ acting only on *polynomial* functions; or alternatively as a linear map $D\colon \mathcal{C}^{k+1}(I) \to \mathcal{C}^k(I)$ which lowers the degree of differentiability by one.
  - It is good to keep these in mind and to be flexible in one's interpretation of D.
  - (There is a modern answer to the question of what is the "correct" "largest" space of functions on which D can be said to act: this is the space of so-called Schwartz distributions; these are generalizations of functions, which include all of the continuous functions (even the non-differentiable ones!), and include (useful!) exotic entities like the *Dirac delta distribution*, which is the mass-distribution function of a point particle with non-zero mass. We may return to this later.)
- We recall that the set $\mathcal{L}(V)$ of operators on any vector space V itself forms a vector space.
  - Thus we can form new *differential operators* by addition and scalar multiplication. For example $(2D + 3\mathrm{I})y = 2Dy + 3\mathrm{I}y = 2y' + 3y$, where $\mathrm{I} \in \mathcal{L}(V)$ is the *identity operator*; normally we just omit I and write $2D + 3$ in place of $2D + 3\mathrm{I}$.
  - Moreover, we have the bilinear composition map $\circ\colon \mathcal{L}(V) \times \mathcal{L}(V) \to \mathcal{L}(V)$; given $f, g \in \mathcal{L}(V)$, we may just write $g \cdot f$ or gf for $g \circ f$ and write f^2 for $f \circ f$.

  - We can thus obtain further examples by composition: for example $(D^2-1)y = D^2y - 1\cdot y$ and $(D+s)(D+t)y = (D+s)(y'+ty) = D(y'+ty) + s(y'+ty) = y'' + (s+t)y' + sty$. Note that using bilinearity of composition, we can say directly that $(D+s)(D+t) = D^2 + (s+t)D + st$.

- We may thus say that a second-order linear differential equation with constant coefficients is one of the form $(D^2 + aD + b)y = f(x)$, the entity on the left being a *second-order linear differential operator*.

Characteristic equations

- Let us now consider a *homogeneous* second-order linear differential equation $y'' + ay' + b = 0$.
- Motivated by the first-order homogeneous linear equation, let us look for solutions of the form $y = e^{rx}$.
  - (Such a guess about the general form of a solution is often referred to using the German word *Ansatz*, meaning "approach" or "starting point"; so we can say "we make the *Ansatz* $y = e^{rx}$".)
  - We find that this y is a solution if and only if $(r^2 + ar + b)e^{rx} = 0$.
  - The resulting equation $r^2 + ar + b = 0$ is called the characteristic equation associated to the given differential equation.

Example 11.1.3

- $y'' - 3y' + 2y = 0$ has characteristic equation $r^2 - 3r + 2 = 0$.
- Factoring, we find this has roots 1 and 2.
- Thus $y_1(x) = e^x$ and $y_2(x) = e^{2x}$ are both solutions.
- Since the operator $L = D^2 - 3D + 2$ (for which our equation is $Ly = 0$) is *linear*, we have $L(c_1y_1 + c_2y_2) = 0$ for any $c_1, c_2 \in \mathbb{R}$.
- Thus, $y(x) = c_1e^x + c_2e^{2x}$ is a solution for any $c_1, c_2 \in \mathbb{R}$.
- In fact, this is the general solution, as we will learn how to prove next.

12 Mar 5, 11.1B and 11.2A: More second-order equations

12.1 11.1B: Factoring operators

- To find the general solution to a homogeneous second-order linear differential equation $Ly = 0$, we factor L and reduce to solving first-order linear equations.

Example 11.1.4

- We seek all solutions $y\colon \mathbb{R} \to \mathbb{R}$ to $y'' + 5y' + 6 = 0$.
- We have $D^2 + 5D + 6 = (D + 3)(D + 2)$.
- Thus, we are seeking solutions of $(D + 3)(D + 2)y = 0$.
- We see that that the function $u = (D + 2y)$ must solve the first-order linear ODE $(D + 3)u = 0$; but we know the solutions to such an ODE: we must have $(D + 2)y = u = c_1e^{-3x}$ for some $c_1 \in \mathbb{R}$.
- We have reduced to the (non-homogeneous) first-order ODE $y' + 2y = c_1e^{-3x}$; we know how to solve this with exponential multipliers.
- We have $e^{2x}(y' + 2y) = c_1e^{-x}$, hence $\frac{\mathrm{d}}{\mathrm{d}x}(ye^{2x}) = c_1e^{-x}$, hence $ye^{2x} = -c_1e^{-x} + c_2$, hence $y = -c_1e^{-3x} + c_2e^{-2x}$.
- We have thus found the general solution; as expected, it is the same as the one resulting from our Ansatz $y = e^{rx}$.

Theorem 11.1.1

- The above procedure leads to the following general theorem:
  - Given a differential equation $y'' + ay' + b = 0$, if the corresponding characteristic equation $r^2 + ar + b = 0$ has two distinct roots $r_1, r_2 \in \mathbb{R}$, then the general solution $y\colon \mathbb{R} \to \mathbb{R}$ to the equation is $y = c_1e^{r_1x} + c_2e^{r_2x}$.
  - If the characteristic equation has a double root $r_1 = r_2 \in \mathbb{R}$, then the general solution is $y = c_1e^{r_1x} + c_2xe^{r_1x}$.
  - In both cases, for any $x_0, y_0, z_0 \in \mathbb{R}$, the initial value problem

$$\begin{cases} y'' + ay' + b = 0 \\ y(x_0) = y_0 \\ y'(x_0) = z_0 \end{cases}$$

    has a unique solution (i.e., these initial conditions uniquely determine the coefficients c_1, c_2 in the general solution).

- The proof:
  - As above, we can write the equation as $(D - r_1)(D - r_2)y = 0$.
  - We thus have that $y\colon \mathbb{R} \to \mathbb{R}$ is a solution if and only if $(D-r_2)y = c_1e^{r_1x}$, i.e., $y'-r_2y = c_1e^{r_1x}$ for some $c_1 \in \mathbb{R}$.
  - This is equivalent to $\frac{\mathrm{d}}{\mathrm{d}x}(ye^{-r_2x}) = c_1e^{(r_1-r_2)x}$.
  - If $r_1 \neq r_2$, this is equivalent to $ye^{-r_2x} = (r_1 - r_2)^{-1}c_1e^{(r_1-r_2)x} + c_2$ for some $c_2 \in \mathbb{R}$, and hence to $y = c_1e^{r_1x} + c_2e^{r_2x}$ for some $c_1, c_2 \in \mathbb{R}$.

- If $r_1 = r_2$, so that $e^{(r_1-r_2)x} = 1$, the above is equivalent to $ye^{-r_2x} = c_1x + c_2$ for some $c_2 \in \mathbb{R}$, and hence to $y = c_1e^{r_1x} + c_2xe^{r_1x}$ for some $c_1, c_2 \in \mathbb{R}$.
- In the first case, the initial conditions $y(x_0) = y_0$ and $y'(x_0) = z_0$ give

$$\begin{cases} y_0 = c_1e^{r_1x_0} + c_2e^{r_2x_0} \\ z_0 = r_1c_1e^{r_1x_0} + r_2c_2e^{r_2x_0} \end{cases}$$

- In the second case, they give

$$\begin{cases} y_0 = c_1e^{r_1x_0} + c_2x_0e^{r_2x_0} \\ z_0 = r_1c_1e^{r_1x_0} + c_2(1 + r_1x_0)e^{r_2x_0} \end{cases}$$

- In both cases, this is a system of two *linear equations* in c_1 and c_2, and we claim in both cases that this has a unique solution $(c_1, c_2) \in \mathbb{R}^2$.

- In fact, it is not hard to solve these systems of linear equations directly, but it is better to do a little more linear algebra review and recall some general principles.
  - Namely, recall that an inverse of a square matrix $A \in K^{n\times n}$, is a matrix $A^{-1} \in K^{n\times n}$ with the property that $A \cdot A^{-1} = A^{-1} \cdot A = \mathrm{I}_n$. If A has an inverse (or as we say, if it is invertible), it has a *unique* inverse.
  - If A is invertible, then for any $b \in K^n$, the system of equations $A\mathbf{x} = \mathbf{b}$ has the unique solution $\mathbf{x} = A^{-1}\mathbf{b}$.
  - A criterion for invertibility is that A is invertible if and only if its *determinant* $\det A$ is non-zero.
  - We will recall more about determinants later; for now, we recall that for a 2×2-matrix $A \in K^{2\times 2}$, the determinant is given by $\det A = A_{11}A_{22} - A_{12}A_{21}$.

- Returning to the above case, the determinants of the matrices appearing in those systems of linear equations are in the first case $e^{(r_1+r_2)x_0}(r_1 - r_2)$, which is non-zero since we are assuming $r_1 \neq r_2$, and in the second case $e^{(r_1+r_2)x_0}(1 + r_1x_0 - r_1x_0) \neq 0$.

12.2 11.2A: Complex Exponentials

- We have solved the homogeneous second-order linear equation $y'' + ay' + b = 0$ in the case that the characteristic equation $r^2 + ar + b = 0$ has two *real* roots.

- As we know, a general quadratic polynomial may have complex roots; to deal with the corresponding differential equation in this case, we must introduce the *complex exponential function.*

- For this purpose, let us first ask ourselves: what *is* the exponential function, and also: what are the cosine and sine functions?

- As we have seen, one way to characterize the exponential function is as the unique solution $y = e^x$ to the IVP $y' = y;\ y(0) = 1$.
  - We saw how to prove uniqueness *assuming* that we already have the exponential function satisfying $\frac{\mathrm{d}}{\mathrm{d}x}e^x = e^x$ and $e^0 = 1$; but how do we know that such a function exists?
  - One way is to use power series.

- Remember that any power series function $f(x) = \sum_{n=0}^{\infty} a_nx^n$ has a *radius of convergence* $R \in [0, \infty]$, such that the series converges whenever $|x| < R$, and diverges when $|x| > R$ (and this is true even for $x \in \mathbb{C}$!).

- One can see (for example using the comparison test against a geometric series) that the radius of convergence of the series $\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!}$ is ∞. (This series should look familiar: it is the Taylor series of the exponential function; but now, we are using this series to *define* the exponential function!).
- Moreover, one can show that any power series function $f(x) = \sum_{n=0}^{\infty} a_n x^n$ is differentiable within its radius of convergence and its derivative is given term-by-term: $f'(x) = \sum_{n=0}^{\infty} n a_n x^{n-1}$.
- It follows that $\exp'(x) = \exp(x)$, and we clearly have $\exp(0) = 1$, so this proves the existence of a solution to the IVP defining e^x.

- Note that the addition law $e^{a+b} = e^a \cdot e^b$ follows from the characterization of e^x as the unique solution to $y' = y;\ y(0) = 1$.
  - Indeed, if we set $y(x) = e^{a+x}/e^a$, then $y' = y$ (by the chain rule) and $y(0) = e^a/e^a = 1$, and hence $y(x) = e^x$, i.e., $e^{a+x} = e^a \cdot e^x$.
  - In particular, this also gives $1 = e^{a-a} = e^a \cdot e^{-a}$ and hence $e^{-a} = 1/e^a$.

13 Mar 10, 11.2A: Complex solutions

13.1 More on complex exponentials

- Last time, we defined the exponential function as $e^z = \sum_{n=0}^{\infty} \frac{z^n}{n!}$.
- A consequence of this is that we can immediately make sense of e^z for *complex* values $z \in \mathbb{C}$, simply by plugging z into the power series.
  - In particular, for $x \in \mathbb{R}$, we find that

$$\begin{aligned} e^{ix} &= 1 + ix - \frac{x^2}{2} - i\frac{x^3}{3!} + \frac{x^4}{4!} + \cdots \\ &= \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{2n!} + i\sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!} \end{aligned}$$

  - We recognize the two terms appearing as the Taylor series for cos and sin.
  - As with the exponential function, let us *define* cos and sin to be given by these two power series.
- It follows immediately from this definition that $\sin' = \cos$ and $\cos' = -\sin$.
  - Soon, we will see that there are in fact *unique* functions satisfying the initial value problems $y'' = -y; y(0) = 0; y'(0) = 1$ and $y'' = -y; y(0) = 1; y'(0) = 0$, respectively. Hence, as with the exponential function, we get a nice condition characterizing cos and sin uniquely, and the above series prove the existence of functions satisfying these conditions.
  - It is also not too hard to show from these definitions that $\cos^2 + \sin^2 = 1$ and that they agree with the geometric definition of the trigonometric functions, i.e., that $\mathbf{p} = (\cos(\theta), \sin(\theta)) \in \mathbb{R}^2$ is the point on the unit circle such that the arc-length from $(1, 0)$ is $\mathbf{p}$, measured counter-clockwise, is θ.
- This way of defining e^x, cos, and sin immediately gives rise to the famous *Euler's formula*:

$$e^{ix} = \cos x + i \sin x.$$

  - This gives a concise way to express points in the plane using polar coordinates: the point with radius r and angle θ is $re^{i\theta}$.
    - As usual with radial coordinates, the angle is not uniquely determined: we always have $re^{i\theta} = re^{i(\theta+2\pi)}$ (and conversely, if $r_1 e^{i\theta_1} = r_2 e^{i\theta_2}$, then $r_1 = r_2$ and $\theta_1 - \theta_2 \in 2\pi\mathbb{Z}$ – the one exception being that $0e^{i\theta} = 0$ for *any* θ)

Some properties of the complex exponential

- A variant (using a bit of complex analysis) of the argument given above to deduce the exponential law $e^{a+b} = e^a \cdot e^b$ for $a, b \in \mathbb{R}$ proves that this holds as well for $a, b \in \mathbb{C}$.
  - From this, we can deduce the addition laws for sin and cos:

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

    hence by comparing real and imaginary parts, we get $\cos(a+b) = \cos a \cos b - \sin a \sin b$ and $\sin(a+b) = \sin a \cos b + \cos a \sin b$.

- This also makes complex numbers easy to multiply when written in polar coordinates: $(r_1 e^{i\theta_1})(r_2 e^{i\theta_2}) = (r_1 r_2)e^{i(\theta_1+\theta_2)}$.
- In particular, this allows us to easily find square roots (and more generally n-th roots): if $z = re^{i\theta}$, then its square roots – i.e., the numbers $w \in \mathbb{C}$ such that $w^2 = z$ – are just $w = \pm\sqrt{r}e^{i\theta/2}$.
  - Indeed, we see that these *are* square roots, and given any other square root w, we have $w^2 - z = (w - \sqrt{r}e^{i\theta/2})(w + \sqrt{r}e^{i\theta/2})$, and hence $w = \pm\sqrt{r}e^{i\theta/2}$.
  - (Regarding the ambiguity of θ: had we written $z = re^{i\theta+2pi}$, we would have gotten the same square roots $w = \pm\sqrt{r}e^{i\theta/2+\pi} = \mp\sqrt{r}e^{i\theta/2}$.
- In particular, if z is a negative real number, then $z = re^{i\pi}$, and we have the familiar imaginary square root $\sqrt{z} = \pm\sqrt{r}e^{i\pi/2} = \pm i\sqrt{r}$.

- Next, recall that the derivative of a function $f\colon \mathbb{R} \to \mathbb{C} = \mathbb{R}^2$ is defined component-wise: if $f(x) = u(x) + iv(x)$, then $f'(x) = u'(x) + iv'(x)$.
  - It follows that
$$\tfrac{\mathrm{d}}{\mathrm{d}x}e^{ix} = \cos' x + i\sin' x = -\sin x + i\cos x = ie^{ix}$$
  - Hence, the *anti-derivative* $\int e^{ix}\,\mathrm{d}x$ of e^{ix} (the unique-up-to-a-constant function $f\colon \mathbb{R} \to \mathbb{C}$ whose derivative is e^{ix}) is $\frac{1}{i}e^{ix} + C = -ie^{ix} + C$, where $C \in \mathbb{C}$ is an arbitrary *complex* constant.
  - More generally, using that $e^{a+ib} = e^a \cdot e^{ib}$, we have that $\frac{\mathrm{d}}{\mathrm{d}x}e^{(a+ib)x} + (a+ib)e^{(a+ib)x}$ and $\int e^{(a+ib)x}\,\mathrm{d}x = \frac{1}{a+ib}e^{(a+ib)x} + C$.

13.2 11.2A: Complex solutions

Example 11.2.1

- Now consider a general homogeneous second order linear equation $Ly = (D^2 + aD + b)y = 0$, with its characteristic equation $r^2 + ar + b = 0$.
  - We are perhaps still mainly interested in solutions $y\colon \mathbb{R} \to \mathbb{R}$, but now we can also try to find all solutions $y\colon \mathbb{R} \to \mathbb{C}$; and in this case, we can also consider equations with coefficients $a, b \in \mathbb{C}$.
- We now factor this polynomial as $(r - r_1)(r - r_2)$, with roots $-a \pm \frac{1}{2}\sqrt{a^2 - 4b}$ (where this square root is now possibly complex).
- We can thus factor the differential operator as $L = (D - r_1)(D - r_2)$.
  - We will see that the above exponential multiplier method still works because $D(e^{rx}y) = e^{rx}(D + r)y$ as before, even for $r \in \mathbb{C}$.
- As a first example, consider $y'' + y = 0$, i.e., $(D^2 + 1)y = 0$, i.e., $(D - i)(D + i)y = 0$.
  - We would like to conclude from this that $(D + i)y = c_1 e^{ix}$ for some $c_1 \in \mathbb{C}$.
  - That is, we would like to say that $u = Ce^{ix}$ is the *general* solution $u\colon \mathbb{R} \to \mathbb{C}$ to the differential equation $u' = iu$.
  - Indeed, the usual method works: given any solution u, we have $\frac{\mathrm{d}}{\mathrm{d}x}(ue^{-ix}) = u'e^{-ix} - iue^{-ix} = 0$, hence $ue^{-ix} = C$ for some constant $C \in \mathbb{C}$, and hence $u = Ce^{ix}$.
- It now remains to solve $y' + iy = c_1 e^{ix}$, and for this, we again use exponential multipliers.

- This equation is equivalent to $e^{ix}(y' + iy) = c_1 e^{2ix}$ (because we can multiply e^{-ix} to go back!), or in other words $\frac{\mathrm{d}}{\mathrm{d}x}(ye^{ix}) = c_1 e^{2ix}$.
  - Note that this use of the product rule is legitimate: it is just the ordinary product rule applied to each of the two components of the function $\mathbb{R} \to \mathbb{C}$ given by $x \mapsto y(x)e^{ix}$.
- This is equivalent to $ye^{ix} = \frac{1}{2}c_1 e^{2ix} + c_2$ for some $c_2 \in \mathbb{C}$ and hence to

$$y = c_1 e^{ix} + c_2 e^{-ix}$$

for some $c_1, c_2 \in \mathbb{C}$; thus, this is the general solution.

- We can rewrite this as

$$\begin{aligned} y &= c_1(\cos x + i \sin x) + c_2(\cos x - i \sin x) \\ &= (c_1 + c_2)\cos x + (c_1 - c_2)i \sin x \\ &= d_1 \cos x + d_2 \sin x, \end{aligned}$$

where we have set $d_1 = c_1 + c_2$ and $d_2 = i(c_1 - c_2)$.

- Since we can solve for c_1 and c_2 in terms of d_1 and d_2, this formula also expresses the general solution.
- It is the general solution we might have expected for $y'' = -y$, except that now d_1 and d_2 may be *complex*.
- The general *real* solution is obtained by restricting to the case $d_1, d_2 \in \mathbb{R}$.

Theorem 11.2.3

- The above example shows how the proof of Theorem 11.1.1 above can be adapted to the complex context, to yield the following statement:
- Given any $a, b \in \mathbb{C}$, the differential equation $y'' + ay' + by = 0$ has the general solution $y \colon \mathbb{R} \to \mathbb{C}$ given by

$$\begin{cases} y = c_1 e^{r_1 x} + c_2 e^{r_2 x}, & r_1 \neq r_2 \\ y = c_1 x e^{r_1 x} + c_2 e^{r_2 x}, & r_1 = r_2, \end{cases}$$

with $c_1, c_2 \in C$, where $r_1, r_2 \in \mathbb{C}$ are the roots of $r^2 + ar + b = 0$.

- In the case where $a, b \in \mathbb{R}$ and $a^2 - 4b < 0$, so that $r_1 = \alpha + i\beta$ and $r_2 = \alpha - i\beta$ for some $\alpha, \beta \in \mathbb{R}$, the general solution can also be written

$$y = c_1 e^{\alpha x} \cos \beta x + c_2 e^{\alpha x} \sin \beta x,$$

where $c_1, c_2 \in \mathbb{C}$. In this case the *real* solutions $y \colon \mathbb{R} \to \mathbb{R}$ are precisely those with $c_1, c_2 \in \mathbb{R}$.

- Moreover, the initial conditions $y(x_0) = y_0$ and $y'(x_0) = z_0$, for any $x_0 \in \mathbb{R}$ and $y_0, z_0 \in \mathbb{C}$ can always be satisfied by a unique choice of c_1 and c_2.

Example 11.2.2

- A generalization of the equation $y'' + y = 0$ from Example 11.2.1 is the equation $y'' + \omega^2 y = 0$, where $\omega \in \mathbb{R} - \{0\}$; this is called the *harmonic oscillator equation* with *angular frequency* ω.
- This has characteristic equation $r^2 + \omega^2 = 0$ with roots $r = \pm i\omega$, and hence general solution

$$y = c_1 e^{i\omega x} + c_2 e^{-i\omega x} = d_1 \cos \omega x + d_2 \sin \omega x.$$

- Such functions are called *harmonic oscillators*, and arise frequently in physics.

Example 11.2.3

- We summarize the form of the general solution to $y'' + ay' + by = 0$ with $a, b \in \mathbb{R}$.
- This has characteristic equation $r^2 + ar + b =$ with roots $r_1, r_2 = -a/2 \pm \sqrt{a^2 - 4b}/2$.
- We now consider three cases, based on the sign of the discriminant $a^2 - 4b$:
- If $a^2 - 4b > 0$, then r_1, r_2 are real and distinct, and the general solution is $y = c_1 e^{r_1 x} + c_2 e^{r_2 x}$ with $c_1, c_2 \in \mathbb{R}$.
- If $a^2 - 4b < 0$, then r_1, r_2 are imaginary and distinct, and the general solution is $y = c_1 e^{\alpha x} \cos \beta x + c_2 e^{\alpha x} \sin \beta x$ with $c_1, c_2 \in \mathbb{R}$, where $\alpha = -a/2$, and $\beta = \sqrt{a^2 - 4b}/2$.
- If $a^2 - 4b = 0$ represents the dividing line between the above oscillatory and non-oscillatory behaviour; the general solution is $y = c_1 x e^{r_1 x} + c_2 e^{r_2 x}$.

14 Mar 24, 11.2B: Higher-order equations

14.1 11.2B: Higher-order equations

Theorem 11.2.4

- The same idea that went in to solving linear homogeneous second order equations with constant coefficients also allows us to solve such equations of arbitrary order.
- Thus, given any $a_0, \dots, a_{n-1} \in \mathbb{C}$, we consider the differential equation

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

  which has a **characteristic equation**

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

  with (by the fundamental theorem of algebra!) n roots (possibly with repetitions) $r_1, \dots, r_n$, so that

  $$r^n + a_{n-1}r^{n-1} + \cdots + a_1 r + a_0 = (r - r_1)(r - r_2)\cdots(r - r_n)$$

  and hence

  $$L = (D - r_1)(D - r_2)\cdots(D - r_n).$$

- **Warning**: recall that, for $n \geq 5$, there is no general formula for finding the roots of a degree n polynomial (and even for $n = 3$ and $n = 4$, the formulas are quite complicated), so the theorem we are about to state gives us the solutions to this differential equation, but *only if we are able to find the roots of the characteristic equation.*
  - (As an aside, if you do not know how to prove the fundamental theorem of algebra, *you should be curious as to how it's proven*!)
- The resulting theorem statement is then: if the roots r_k are all *distinct*, then the equation $Ly = 0$ has the general solution $y \colon \mathbb{R} \to \mathbb{C}$ given by

  $$y = c_1 e^{r_1 x} + \cdots + c_n e^{r_n x}$$

  with $c_1, \cdots, c_n \in \mathbb{C}$.
  - If there are repetitions among the roots, so that, say $\{r_1, \dots, r_n\} = \{s_1, \dots, s_m\}$, with s_k appearing d_k times, then the general solution is

    $$y = \sum_{k=1}^{m} \sum_{j=0}^{d_k - 1} c_{k,j} x^j e^{s_k x},$$

    with $c_{k,j} \in \mathbb{C}$ for $1 \leq k \leq m$ and $0 \leq j < d_k$. (Note that this actually includes the previous case, which arises when $d_k = 1$ for all k, so that $s_k = r_k$).
  - Moreover, for any $y_0, \dots, y_{n-1} \in \mathbb{C}$ and $x_0 \in \mathbb{R}$, there are unique values of $c_1, \dots, c_n$ (or of $c_{k,j}$ for $1 \leq k \leq m$ and $0 \leq j < d_k$) so that y satisfies the initial conditions $y(x_0) = y_0; \dots; y^{(n-1)}(x_0) = y_{n-1}$.
- To prove this theorem, we must (i) verify that the y given above is indeed a solution, (ii) prove that it is the *general* solution, i.e., that every solution has this form, and (iii) verify the claim about initial conditions. Step (i) is fairly straightforward. As indicated above, step (ii) proceeds by induction, with each induction step consisting in solving a first-order (possibly non-homogeneous) linear equation. Step (iii) involves, as in the second-order case, checking that a certain matrix has non-zero determinant.

- To verify that y as above is a solution, it suffices by linearity of L to show that $x^j e^{s_k x}$ is a solution for each $1 \leq k \leq m$ and $0 \leq j < d_k$.
  - The relevant general fact is that $(D-r)^b(x^a e^{rx}) = 0$ whenever $0 \leq a < b$, which we prove by induction on b.
    * The base case $b = 1$ we already know.
    * For the induction step, assume $b > 0$ and that the claim is already known for $b-1$.
    * We then have $(D-r)^b(x^a e^{rx}) = (D-r)^{b-1}(D-r)(x^a e^{rx})$.
    * If $a = 0$, then $(D-r)(x^a e^{rx}) = 0$, and we are done.
    * If $a > 0$, then
    $$(D-r)(x^a e^{rx}) = (ax^{a-1}e^{rx} + rx^a e^{rx}) - r(x^a e^{rx}) = ax^{a-1}e^{rx},$$
    and so since $a - 1 \leq b - 1$, we are done by the induction hypothesis.
  - Now to show that $L(x^j e^{s_k x}) = 0$ for $0 < j < d_k$, we use that the $(D - s_k)$ all commute with each other to write L as a product $\widetilde{L}(D-s_k)^{d_k}$ (where $\widetilde{L}$ is the product of all the $(D-s_{k'})^{d_{k'}}$ for $k \neq k'$).
    * We then have $L(x^j e^{s_k x}) = \widetilde{L}(D-s_k)^{d_k}(x^j e^{s_k x}) = 0$.
- Let us now prove by induction that y as given above is the general solution.
  - As a base case, we can take the case $n = 2$, which we already know (or alternatively, we can take the trivial case $n = 0$, so that the equation is $y = 0$, with unique solution $y = 0$).
  - For the induction step, assume that $n > 0$ and that the theorem holds for equations of degree $n - 1$.
  - Now let y be a solution, so that $Ly = 0$.
  - We set $u = (D - r_n)y$, so that we can write the equation as
  $$(D-r_1)(D-r_2)\cdots(D-r_{n-1})u = 0.$$
  - By the induction hypothesis, we thus have that u is a linear combination
  $$(D-r_n)y = \sum_{k=1}^{m-1}\sum_{j=0}^{d_k-1} c_{k,j}x^j e^{s_k x}$$
  for some $c_{k,j} \in \mathbb{C}$ if $r_n \notin \{r_1, \ldots, r_{n-1}\}$ and
  $$(D-r_n)y = \sum_{k=1}^{m-1}\sum_{j=0}^{d_k-1} c_{k,j}x^j e^{s_k x} + \sum_{j=0}^{d_m-2} c_{m,j}x^j e^{s_m x}$$
  if $r_n \in \{r_1, \ldots, r_{n-1}\}$ (and where $r_n = s_m$).
  - Using the usual exponential multiplier, these are equivalent to
  $$\frac{\mathrm{d}}{\mathrm{d}x}(e^{-r_n x}y) = \sum_{k=1}^{m-1}\sum_{j=0}^{d_k-1} c_{k,j}x^j e^{(s_k-r_n)x}$$
  and
  $$\frac{\mathrm{d}}{\mathrm{d}x}(e^{-r_n x}y) = \sum_{k=1}^{m-1}\sum_{j=0}^{d_k-1} c_{k,j}x^j e^{(s_k-r_n)x} + \sum_{j=0}^{d_m-2} c_{m,j}x^j,$$
  respectively.

- Using that for any integer N and $R \in \mathbb{C}$, $\int x^N e^{Rx}\,\mathrm{d}x$ is (a constant plus) a linear combination of $x^M e^{Rx}$ for $0 \leq M \leq N$ (as one verifies by induction using integration by parts), we conclude that $e^{-r_n x}y$ is,
  - in the first case, a constant c_n plus linear combination of $x^j e^{(s_k - r_n)x}$ for $1 \leq k \leq m-1$ and $0 \leq j \leq d_k - 1$
  - in the second case, a constant c_n plus a linear combination of $x^j e^{(s_k - r_n)x}$ for $1 \leq k \leq m-1$ and $0 \leq j \leq d_k - 1$, plus a linear combination of x^{j+1} for $0 \leq j \leq d_m - 2$.
- We thus conclude, as desired, that y is:
  - in the first case, a linear combination of $e^{r_n x} = e^{s_m x}$ and of $x^j e^{s_k}$ for $1 \leq k \leq m-1$ and $0 \leq j \leq d_k - 1$
  - in the second case, a linear combination of $x^j e^{s_k}$ for $1 \leq k \leq m$ and $0 \leq j \leq d_k - 1$.

- Finally, let us see that for each $y_0, \ldots, y_{n-1} \in \mathbb{C}$ and $x_0 \in \mathbb{R}$, there are unique coefficients in the general solution for which y satisfies the initial conditions $y(x_0) = y_0;\ \ldots;\ y^{(n-1)}(x_0) = y_{n-1}$.
  - It is helpful first to reduce to the case $x_0 = 0$.
    - Thus, for any $y\colon \mathbb{R} \to \mathbb{C}$, consider the function $z(x) = y(x - x_0)$; we have $z^{(l)}(x) = y^{(l)}(x_0)$ for all l and hence (because of homogeneity!) z we have $Lz = 0$ if and only if $Ly = 0$.
    - The assignment $y \mapsto z$ is clearly a bijection (with inverse taking z to the functions y defined by $y(x) = z(x + x_0)$).
    - Hence, given any $y_0, \ldots, y_{n-1}$, we obtain a bijection between solutions z satisfying $z(0) = y_0, \ldots, z^{(n-1)}(0) = y_{n-1}$ and solutions y satisfying $y(x_0) = y_0, \ldots, y^{(n-1)}(x_0) = y_{n-1}$.
    - Thus, if there is a unique such solution z, there is a unique such solution y.
  - Next, let's first consider the case in which all the roots r_i are distinct.
    - If $y = \sum_{k=1}^n c_k e^{r_k x}$ is the general solution, we then have $y^{(l)} = \sum_{k=1}^n c_k r_k^l e^{r_k x}$ and hence $y^{(l)}(0) = \sum_{k=1}^n c_k r_k^l$ for all l.
    - The initial conditions thus give a system of linear equations

      $$y_l = \sum_{k=1}^n r_k^l c_k$$

      for $l = 0, \ldots, n-1$ in the variables $c_1, \ldots, c_n$.
    - We want to show that this has a unique solution for any $y_0, \ldots, y_{n-1}$; in other words, we are asking about the invertibility of the matrix $(r_k^l)_{1 \leq k \leq n, 0 \leq l \leq n-1}$ (assuming the r_k are distinct!).
    - This is one of the world's most famous matrices, and is called the Vandermonde matrix, after a 18th century French mathematician.
    - In fact, it is known to have determinant $\prod_{1 \leq j < k \leq n}(x_k - x_j)$, which is thus non-zero since the x_i are distinct; however, we can also see the invertibility directly, by showing that it represents an invertible linear map.
    - Namely, we consider the linear map $f\colon \mathbb{C}[x]_{x<n} \to \mathbb{C}^n$ given by $f(p) = (p(x_1), \ldots, p(x_n))$; with respect to the standard (monomial) basis of $\mathbb{C}[x]_{x<n}$ and the standard basis of $\mathbb{C}^n$, this is indeed represented by the Vandermonde matrix (or perhaps by its transpose).
    - Now for p to be in the kernel of this map means that the degree $< n$ polynomial p has n distinct roots $x_1, \ldots, x_n$; but by the fundamental theorem of algebra, this can only happen if $p = 0$. Hence f has trivial kernel (and is hence invertible by the rank-nullity theorem).
  - Finally, we consider the general case (where there may be multiple roots).

* As a first lemma, we need to know the derivatives of $y(x) = x^j e^{rx}$ for $r \in \mathbb{C}$ and $N \geq 0$.
  - More generally, we consider $f(x) = g(x)h(x)$ for any two functions f and g.
  - It's then easy to see by induction that $f^{(l)}(x) = \sum_{m=0}^{l} \binom{l}{m} g^{(m)}(x) h^{(l-m)}(x)$.
  - Applying this with $f(x) = g(x)h(x) = x^j e^{rx}$ we have that $g^{(m)}(0) = \delta_{jm} j!$ and $h^{(m-j)}(0) = r^{m-j} e^{r0} = r^{m-j}$, and hence

$$f^{(l)}(0) = \sum_{m=0}^{l} \binom{l}{m} \delta_{jm} j! r^{m-j} = \begin{cases} \binom{l}{j} j! r^{l-j} = \frac{l!}{(l-j)!} r^{l-j} & j \leq l \\ 0 & j > l \end{cases}$$

* Thus, taking $y = \sum_{k=1}^{m} \sum_{j=0}^{d_k - 1} c_{k,j} x^j e^{s_k x}$ as above, we conclude that

$$y^{(l)} = \sum_{k=1}^{m} \sum_{j=0}^{\max(d_k - 1, l)} c_{k,j} \frac{l!}{(l-j)!} s_k^{l-j}.$$

  - As before, we want to consider this as a linear combination of the $c_{k,j}$. If we just look at the coefficients of $c_{k,j}$ for a fixed value of k, we obtain the following sequence – let us call it C_k^l:

$$C_k^l = \Big(\frac{l!}{(l-0)!} s_k^l, \frac{l!}{(l-1)!} s_k^{l-1}, \dots, \frac{l!}{0!} s_k^0, 0, \dots 0 \Big).$$

  - If we consider the coefficients of all of the $c_{k,j}$, we thus obtain the concatenation of these sequence $C_1^l \dots C_m^l$.
  - Hence, finally, the matrix A corresponding to the system of equations $y_l = y^{(l)}(0)$ (with $l = 0, \dots, n-1$), which we have to show is invertible, looks as follows:

$$A = \begin{bmatrix} C_1^0 & C_2^0 & \cdots & C_m^0 \\ & & \vdots & \\ C_1^{n-1} & C_2^{n-1} & \cdots & C_m^{n-1} \end{bmatrix},$$

    where it is to be remembered that each C_k^l is itself a sequence of entries (arranged horizontally).

* We now claim again that the resulting matrix A (or rather its transpose) represents an invertible linear map $\mathbb{C}[x]_{<n} \to \mathbb{C}^n$.
  - Namely, we consider the map $f \colon \mathbb{C}[x]_{<n} \to \mathbb{C}^n$ defined by

$$f(p) = (p(s_1), p'(s_1), \dots, p^{d_1 - 1}(s_1), \cdots, p(s_m), p'(s_m), \dots, p^{d_m - 1}(s_m)).$$

  - We see that if we let p be the l-th basis element x^l of the standard monomial basis of $\mathbb{C}[x]_{<n}$, then the sequence $(p(s_k), p'(s_k), \dots, p^{d_1 - 1}(s_k))$ is precisely the above sequence C_k^l. Hence, the vector $f(p) \in \mathbb{C}^n$ is exactly the l-th row of the matrix A, so that $A^\top$ represents f as claimed.
  - It remains to see that f is an invertible linear transformation; again, it suffices to show that it has a trivial kernel.
  - But if $f(p) = \mathbf{0}$, that means precisely that the polynomial p vanishes at each s_k with multiplicity d_k. (This is a general fact: a polynomial q of order m has a root of multiplicity $d \leq m$ at r if and only if $q(r) = \cdots = q^{(d-1)}(r) = 0$.) Since $\sum_{k=1}^{m} d_k = n$, this means that the degree $< n$ polynomial p has n roots, counted with multiplicity, hence must be 0.

- Let us prove the parenthetical general fact. Note that "q has a root of multiplicity d at r" means precisely that $(x-r)^d$ is a factor of $q(x)$, i.e., $q(x) = (x-r)^d q_d(x)$ for some polynomial q_d.
- The proof is by induction on d. The base case $d = 1$ is the familiar fact that $q(r) = 0$ if and only if $(x-r)$ is a factor of q, and we omit the proof (which uses polynomial division).
- For the induction step, suppose $d > 1$, and write $q(x) = (x-r)q_1(x)$. We wish to show that $q(r) = q'(r) = \cdots = q^{(d)}(r) = 0$ if and only if $(x-r)^d$ divides $q(x)$, i.e., if and only if $(x-r)^{d-1}$ divides q_1.
- Now it is easy to show by induction on l that $q^l(x) = lq_1^{(l-1)}(x) + (x-r)q_1^l(x)$ for $l > 0$.
- It follows that $q^l(a) = 0$ if and only if $q_1^{l-1}(a) = 0$ for all $l > 0$.
- Hence $q'(r) = \cdots = q^{(d)}(r) = 0$ if and only if $q_1(r) = \cdots = q_1^{(d-1)}(r) = 0$, which by the induction hypothesis is equivalent to $(x-r)^{d-1}$ dividing $q_1(x)$, as desired.

Example 11.2.5

- Let's solve $y''' - 4y'' + 4y' = 0$.
- The characteristic equation is $0 = r^3 - 4r^2 + 4r = r(r-2)^2$, with a single root $r_1 = 0$ and a double root $r_2 = 2$.
- Thus the general solution is $y = c_1 + c_2 e^{2x} + c_3 x e^{2x}$.

Example 11.2.6

- We can use Theorem 11.2.4 in reverse, starting with solutions and arriving at a differential equation that has those solutions.
- Thus, a constant-coefficient linear equation $Ly = 0$ of least order having e^x, e^{2x}, and e^{-2x} as solutions is given by $L = (D-1)(D-2)(D+2)$.

15 Mar 26, 11.2C: Independent solutions

15.1 11.2C: Independent solutions

Theorem 11.2.5

- We now describe the solutions to higher order homogeneous linear equations with *real* (constant) coefficients.
- The important fact here is that the roots of a real polynomial $p(r) = r^n + a_{n-1}r^{n-1} + \cdots + a_1 r + a_0$ are all real or come in complex-conjugate pairs $\alpha \pm i\beta$.
  - The reason is that, since the coefficients are real (and since conjugation preserves addition and multiplication), we have $\overline{p(z)} = p(\bar{z})$ and hence $p(z) = 0 \Rightarrow p(\bar{z}) = \bar{0} = 0$.
  - Hence, we can write the roots as $s_1, \ldots, s_m, \alpha_1 \pm i\beta_1, \ldots, \alpha_M \pm i\beta_M$ with multiplicities $d_1, \ldots, d_m, e_1, \ldots, e_M$, so that $\sum_{k=1}^{m} d_k + 2\sum_{k=1}^{M} e_k = n$.
- The theorem statement is now: given an equation $Ly = 0$ with characteristic equation $p(r)$ with roots s_k and $\alpha_k \pm i\beta_k$ as above, the general solution is a linear combination

$$\sum_{k=1}^{m}\sum_{j=1}^{d_k} c_{k,j}x^j e^{s_k x} + \sum_{k=1}^{M}\sum_{j=1}^{e_k}(u_{k,j}x^j e^{\alpha_K x}\cos\beta x + v_{k,j}x^j e^{\alpha_k x}\sin\beta x)$$

  of the functions $x^j e^{s_k x}$, $x^j e^{\alpha_k x}\cos\beta x$, and $x^j e^{\alpha_k x}\sin\beta x$.
  - In the case where all the roots are distinct, this simplifies to

$$\sum_{k=1}^{m} c_k e^{s_k x} + \sum_{k=1}^{M}(u_k e^{\alpha_K x}\cos\beta x + v_k e^{\alpha_k x}\sin\beta x)$$

- As before, this is simply proven by expanding $e^{(\alpha\pm i\beta)x}$ as $e^{\alpha x}(\cos\beta x \pm i\sin\beta x)$ for each such term appearing in the first form of the general solution.
- More specifically, the general *complex* $y\colon \mathbb{R} \to \mathbb{C}$ solution is such a linear combination with $c_{k,j}, u_{k,j}, v_{k,j} \in \mathbb{C}$, and the general *real* solution $y\colon \mathbb{R} \to \mathbb{R}$ is such a linear combination with $c_{k,j}, u_{k,j}, v_{k,j} \in \mathbb{R}$.
  - In one direction, it is clear that *if* the coefficients are all real, then y is real.
  - In the other direction, given any (possibly complex) coefficients, if $y(x) \in \mathbb{R}$ for all x, then $(y(x) + \overline{y(x)}) = 0$ for all x.
  - The function $z(x) = y(x) + \overline{y(x)}$ is a linear combination of the same functions $e^{s_k x}$ and so on, with coefficients $c_{k,j} + \bar{c}_{k,j}$, $u_{k,j} + \bar{u}_{k,j}$, and $v_{k,j} + \bar{v}_{k,j}$.
  - Since $z = 0$, we have in particular that $z(x_0) = z'(x_0) = \cdots = z^{(n-1)}(x_0) = 0$ for any $x_0 \in \mathbb{R}$; and we obviously have $Lz = 0$.
  - Hence, by the uniqueness of the coefficients satisfying given initial conditions, we conclude that $c_{k,j} + \bar{c}_{k,j}$, $u_{k,j} + \bar{u}_{k,j}$, and $v_{k,j} + \bar{v}_{k,j}$ are all 0, and hence that $c_{k,j}$, $u_{k,j}$, $v_{k,j}$ are all real, as desired.

Corollary 11.2.6

- The argument we just gave to prove that the real solutions are exactly those with real coefficients has another nice application: it allows us to prove that the functions $x^j e^{s_k x}$ in the general solution to $Ly = 0$ are *linearly independent* (when considered as elements in the vector space $\mathbb{R}^{\mathbb{R}}$ – or equivalently, in one of its subspaces such as $\mathcal{C}^0(\mathbb{R})$ or $\mathcal{C}^\infty(\mathbb{R})$).
  - Similarly, the functions $x^j e^{s_k x}$, $x^j e^{\alpha_k x} \cos \beta_k x$, and $x^j e^{\alpha_k x} \sin \beta_k x$ are all linearly independent.
- Indeed, if we have some linear combination $y = \sum_{k=1}^{m} \sum_{j=0}^{d_k - 1} c_{k,j} x^j e^{s_k x}$ which is equal to 0, then it satisfies the initial conditions $y(x_0) = y'(x_0) = \cdots = y^{(n-1)}(x_0) = 0$, and hence we must have $c_{k,j} = 0$ for all k, j by the uniqueness of coefficients acidifying given initial conditions.
- Thus, we may say that these functions form a **basis** of the space of solutions to the equation $Ly = 0$ (or in other words, of $\ker L$).

Example 11.2.7

- The equation $y^{(4)} - y = 0$ has characteristic equation $r^4 - 1 = 0$, with roots $1, -1, i, -i$.
- Thus, the general solution is a linear combination of $e^{\pm 1}$ and $e^{\pm i}$, or equivalently, of $e^{\pm 1}$, $\sin x$, and $\cos x$.
- Suppose we impose the initial conditions $y(0) = 0, y'(0) = 1, y''(0) = 2, y'''(0) = 1$.
  - Considering the general solution

    $$y = c_1 e^x + c_2 e^{-x} + c_3 \cos x + c_4 \sin x,$$

    the initial conditions give the equations

    $$\begin{aligned} c_1 + c_2 + c_3 &= 0 \\ c_1 - c_2 + c_4 &= 1 \\ c_1 + c_2 - c_3 &= 2 \\ c_1 - c_2 - c_4 &= 1 \end{aligned}$$

  - We know on principle that the matrix of coefficients of this linear system is invertible, but in this case, it is easy to straightforwardly solve the system.
  - The result is $c_1 = 1$, $c_2 = 0$ $c_3 = -1$, and $c_4 = 0$; thus the unique solution to this IVP is $e^x - \cos x$.

16 Mar 31, 11.3: Nonhomogeneous equations

16.1 11.3A: Superposition

- We now consider non-homogeneous linear equations $Ly = f$, where L is a linear operator $L = D^n + a_{n-1}D^{n-1} + \cdots + a_1 D + a_0$ as before and f is a function.
  - Of course, the homogeneous equations are just the special case when $f = \mathbf{0}$.

Theorem 11.3.1

- There is a basic relationship between homogeneous and non-homogeneous equations, which is very general, and which we know from linear algebra:
  - Given any linear function $f\colon V \to W$ between vector spaces V and W, we can consider the equation $f(\mathbf{x}) = \mathbf{w}$ for some fixed $\mathbf{w} \in W$.
  - Given any **particular solution** $\mathbf{v}_\text{p}$ (so that $f(\mathbf{v}_\text{p}) = \mathbf{w}$, and any $\mathbf{v}_0 \in \ker f$ (i.e., a solution to the homogeneous equation $f(\mathbf{x}) = \mathbf{0}$), we have by linearity that $f(\mathbf{v}_\text{p} + \mathbf{v}_0) = f(\mathbf{v}_\text{p}) = \mathbf{w}$, so $\mathbf{v}_\text{p} + \mathbf{v}_0$ is also a solution.
  - Conversely, given *any* other solution $\mathbf{v}$ (so that $f(\mathbf{v}) = \mathbf{w}$), if we set $\mathbf{v}_0 = \mathbf{v} - \mathbf{v}_\text{p}$, then we have (again by linearity) $f(\mathbf{v}_0) = f(\mathbf{v}) - f(\mathbf{v}_\text{p}) = \mathbf{w} - \mathbf{w} = \mathbf{0}$, so $\mathbf{v}_0 \in \ker f$ and $\mathbf{v} = \mathbf{v}_\text{p} + \mathbf{v}_0$.
  - The conclusion is that *the set of solutions to* $f(\mathbf{x}) = \mathbf{w}$ *is precisely*

$$\{\mathbf{v}_\text{p} + \mathbf{v}_0 \mid \mathbf{v}_0 \in \ker f\}.$$

- In particular, we may apply this to the linear operator L, which is a linear map $L\colon \mathcal{C}^\infty(\mathbb{R}) \to \mathcal{C}^\infty(\mathbb{R})$.
  - The conclusion is that given any particular solution y_p to the inhomogeneous equation $Ly = f$, the general solution will be given by $y = y_\text{p} + y_\text{h}$, where y_h is an arbitrary solution to the associated homogeneous equation $Ly = 0$.

Example 11.3.1

- Consider $y'' + 2y' + y = e^{3x}$.
- The characteristic polynomial is $(r+1)^2$, with the double root $r_1 = r_2 = -1$.
- Thus, we know the general solution to the *homogeneous* equation is $y_\text{h} = c_1 e^{-x} + c_2 x e^{-x}$.
  - Thus, if we can find a *particular* solution y_p to the inhomogeneous equation, we will obtain the general solution as $y = y_\text{p} + y_\text{h} = y_\text{p} + c_1 e^{-x} + c_2 x e^{-x}$.
- Moreover, we can obtain a particular solution by our usual method of iteratively solving (inhomogeneous) first-order equations.
  - Writing $(D+1)^2 y_\text{p} = e^{3x}$ and setting $u = (D+1)y_\text{p}$, we have $u' + u = (D+1)u = e^{3x}$.
  - Introducing the usual exponential multiplier, this is equivalent to $D(e^x u) = e^{4x}$, hence we may take $e^x u = \frac{1}{4}e^{4x}$ (we may ignore the integration constant because we are just looking for *one* solution), and hence $u = \frac{1}{4}e^{3x}$.

- We thus have $y_p' + y_p = (D+1)y_p = \frac{1}{4}e^{3x}$.
- Proceeding exactly as before, we conclude that we may take $y_p = \frac{1}{16}e^{3x}$.

- We conclude that the general solution is

$$y = \frac{1}{16}e^{3x} + c_1 e^{-x} + c_2 x e^{-x}.$$

- (Note that we could also have arrived at this directly – instead of using the general fact that the general solution is a particular solution plus a homogeneous solution – if in the above derivation, we had kept track of the constants of integration.)

Example 11.3.2 ("superposition principle")

- Suppose we have an inhomogeneous equation $Ly = f$, and we are able to write f in the form $f = a_1 f_1 + a_2 f_2$.
- If we want to find a particular solution, it suffices by the linearity of L to find y_1, y_2 with $Ly_1 = f_1$ and $Ly_2 = f_2$, for then setting $y_p = a_1 y_1 + a_2 y_2$, we have $Ly_p = a_1 Ly_1 + a_2 Ly_2 = a_1 f_1 + a_2 f_2$.
- This (as well as other related phenomena) is sometimes called the **superposition principle**, because y_p is a "superposition" of y_1 and y_2.

- For example, suppose we want to solve $y'' + 2y' + y = e^{3x} + 1$.
- We already saw that $y_1 = \frac{1}{16}e^{3x}$ is a solution to $y'' + 2y' + y = e^{3x}$.
- So using the superposition principle, it remains to find a particular solution to $y'' + 2y' + y = 1$.
- We can do this using exponential multipliers as before – or maybe we'll get lucky and notice that there is an obvious solution $y_2 = 1$.
- Hence $y_p = y_1 + y_2 = \frac{1}{16}e^{3x} + 1$ is a particular solution to the equation, and thus $y = \frac{1}{16}e^{3x} + 1 + c_1 e^{-x} + c_2 x e^{-x}$ is the general solution.

- Similarly, suppose we wanted to solve $y'' + 2y' + y = e^{3x} + e^{-x}$.
- Again, we only need to find a particular solution y_2 to $y'' + 2y' + y = e^{-x}$.
- This time, there is not such an obvious solution, but if we use exponential multipliers again, we find that $y_2 = \frac{1}{2}x^2 e^{-x}$ is a solution.
- Hence, the general solution to the equation is $y = \frac{1}{16}e^{3x} + \frac{1}{2}x^2 e^{-x} + c_1 e^{-x} + c_2 x e^{-x}$.

16.2 11.3B: Undetermined coefficients

- In the previous example, in one case we had to carry out the somewhat laborious exponential multiplier method to find a special solution, and in the other case, we got lucky and were able to guess one. We now introduce a method which would have given us a shortcut in both cases.

- The important circumstance is that in both cases, the inhomogeneous term f in $Ly = f$ (namely 1 and e^{-x}, respectively) was *itself* the solution to some *homogeneous* equation $My = 0$ (namely with $M = D$ and $M = D + 1$, respectively).

- It follows that $M(L(y)) = M(f) = 0$.
- But now ML is *itself* some linear differential operator, and thus we know how find the general solution to this homogeneous equation.
- Thus the sought-after particular solution y_p to our original equation is some solution to this homogeneous equation $MLy = 0$, and it just remains to determine the coefficients in the general solution that we need to choose.
- In general, this method, where a solution to a problem is first established as a certain linear combination, the coefficients of which are then determined, is called the **method of undetermined coefficients**.

Example 11.3.1 again

- We again seek a solution y_p to $(D+1)^2 y = e^{-x}$.
  - Since e^{-x} satisfies $(D+1)e^{-x} = 0$, we see that y_p must satisfy $(D+1)^3 y_\text{p} = 0$, and hence be of the form $y_\text{p} = c_1 e^{-x} + c_2 x e^{-x} + c_3 x^2 e^{-x}$.
  - Moreover, we see that the first two terms are actually solutions to the *homogeneous* equation $(D+1)^2 y = 0$, so we may omit them without changing whether y_p is a particular solution.
  - Plugging the remaining term $y_\text{p} = c_3 x^2 e^{-x}$ into the original equation, we obtain
    $$e^{-x} = y_\text{p} + 2y_\text{p}' + y_\text{p}'' = c_3(x^2 + 2(2x - x^2) + (2 - 4x + x^2))e^{-x} = 2c_3 e^{-x}$$
    and hence $c_3 = 1/2$.
- Similarly, if we seek a solution y_p to $(D+1)^2 y = 1$, we see that it must be a solution to $D(D+1)^2 y = 0$, and hence be of the form $y_\text{p} = c_1 + c_2 e^{-x} + c_3 x e^{-x}$.
  - Again, we can ignore the last two terms since they are solutions to the homogeneous equation.
  - We are left with $y_\text{p} = c_1$.
  - We now plug this into the original equation and obtain
    $$1 = y_\text{p} + 2y_\text{p}' + y_\text{p}'' = c_1.$$
- Note in both of these cases the substantial simplification resulting from throwing away the terms which are solutions to the associated homogeneous equation. Without this simplification, this method wouldn't really be more efficient than just solving the equation directly using exponential multipliers.

17 Apr 2, 11.3C: Variation of parameters

17.1 11.3C: Variation of parameters

- We next discuss **variation of parameters**, a further method for finding a particular solution to a non-homogeneous equation, which has the advantage of also working for equations with *non-constant* coefficients.

- This method relies on the following *Ansatz*: given a non-homogeneous equation $Ly = y'' + a(x)y' + b(x)y = f(x)$, and supposing we have solutions y_1, y_2 to the associated *homogeneous* equation $Ly = 0$, we look for a particular solution to the non-homogeneous equation of the form $y_p(x) = u_1(x)y_1(x) + u_2(x)y_2(x)$ for some functions $u_1(x), u_2(x)$.

  - (The name comes from the fact that the coefficients, or *parameters*, in the linear combination of y_1 and y_2 are allowed to depend on x, that is, to *vary*.)

  - If we plug y_p into the original equation and rearrange the terms, we obtain

$$\begin{aligned} f = &(y_1'' + ay_1' + by_1)u_1 + (y_2'' + ay_2' + by_2)u_2 \\ &+ (y_1u_1' + y_2u_2')' + a(y_1u_1' + y_2u_2') \\ &+ (y_1'u_1' + y_2'u_2') \end{aligned}$$

  - Here, the first line vanishes since y_1, y_2 are assumed to be solutions to the homogeneous equation; the second line will vanish if we assume $y_1u_1' + y_2u_2' = 0$; and hence the equation will be satisfied as soon as $f = y_1'u_1' + y_2'u_2'$.

- This gives us a pair of linear equations in u_1' and u_2' (with non-constant coefficients!)

$$\begin{aligned} 0 &= y_1u_1' + y_2u_2' \\ f &= y_1'u_1' + y_2'u_2' \end{aligned}$$

  - This will have a unique solution as long as the determinant $y_1(x)y_2'(x) - y_2(x)y_1'(x)$ is non-zero (for all x!); this is called the Wronskian determinant of y_1 and y_2, and is denoted $w(x)$.

    * (This has a superficial similarity to the Vandermonde determinant – though I don't know what the significance of this is.)

  - (In fact, we will always have $w(x) \neq 0$ as long as y_1 and y_2 are linearly independent solutions; in the case of constant-coefficients, this follows from the uniqueness of solutions with given initial conditions. It is also true with non-constant coefficients, and follows from the uniqueness theorem for higher-order differential equations, which we have not discussed yet.)

  - This shows that the variation of parameters *Ansatz* will work. Before proceeding further with the general method, let us work out a couple of examples of variation of parameters by hand.

Example 11.3.6

- We return to $y'' + 2y' + y = e^{-x}$, where we previously used undetermined coefficients, and again seek a particular solution y_p.

- We now use that $y_1 = e^{-x}$ is a homogeneous solution, and make the *Ansatz* $y_p(x) = u_1(x)e^{-x}$. (According to the general method, we should also have a term with xe^{-x}, but this can be omitted, since any multiple of xe^{-x} is also a multiple of e^{-x}.)

- We have $y_p' = (u_1' - u_1)e^{-x}$ and $y_p'' = (u_1'' - 2u_1' + u_1)e^{-x}$, and hence plugging y_p into the original equation yields

$$e^{-x} = (u_1'' - 2u_1' + u_1)e^{-x} + 2(u_1' - u_1)e^{-x} + u_1 e^{-x} = u_1'' e^{-x}.$$

- This will be satisfied if $u_1'' = 1$, hence we may take $u_1 = \frac{1}{2}x^2$, and hence $y_p = \frac{1}{2}x^2 e^{-x}$, which is the same as we got before.

Example 11.3.7

- We consider $Ly = x^2y'' - 2xy' + 2y = x^4$.
- In the textbook, we are simply told that the associated homogeneous equation has solutions $y_1(x) = x$ and $y_2(x) = x^2$, as one can easily check.
  - To actually arrive at this conclusion, the associated homogeneous equation can be solved by cleverly factoring the operator $L = (x^2D^2 - 2xD + 2)$ (note that this is now a non-trivial task because of the non-constant coefficients!).
  - Let us write $L = (xD - r_1)(xD - r_2)$.
  - We must proceed with caution, and note that for a function $r(x)$, we have $Dr = r' + rD$, because $(Dr)y = D(ry) = r'y + ry' = (r' + rD)y$.
  - Thus, carefully multiplying out the above expression for L, we obtain $L = x^2D^2 + (x - xr_1 - xr_2)D + r_1r_2$.
  - We thus have the equations $x - xr_1 - xr_2 = -2x$, and $r_1r_2 = 2$, with solution $r_1 = 1$ and $r_2 = 2$, and thus $L = (xD - 1)(xD - 2)$.
  - We have reduced the problem of solving $Ly = 0$ to that of solving two first-order linear equations, and when we solve theses, we find we indeed get $y(x) = c_1x + c_2x^2$ as the general solution.
    - (Though we note that this is still a bit tricky since the operators $xD - 1$ and $xD - 2$ are not in "normalized form" $D + g(x)$.)
- We thus make the *Ansatz* $y_p = u_1y_1 + u_2y_2 = u_1x + u_2x^2$.
  - And in fact, as in previous last example, we see that the second term is redundant, so we may make the more specific *Ansatz* $y_p = u_1x$.
  - Thus, $y_p' = u'x + u$ and $y_p'' = u''x + 2u'$
  - Plugging this into the original equation yields

$$x^4 = x^2(u''x + 2u') - 2x(u'x + u) + 2ux = u''x^3.$$

  - Thus, we see that it suffices that u satisfy $u'' = 1$, and hence we can take $u = \frac{1}{6}x^3$.
- We thus have the particular solution $y_p = \frac{1}{6}x^4$, hence the general solution is $y = \frac{1}{6}x^4 + c_1x + c_2x^2$, with c_1 and c_2 *constants*.

11.3D: Green's functions

- In the previous example, it just so happened we were able to get away with just using one of the two independent solutions y_1, y_2 to the homogeneous equation $Ly = 0$ associated to $Ly = y'' + ay' + by = f$; but in general, we need both.
- As we said above, seeking a particular solution of the form $y_\mathrm{p} = u_1 y_1 + u_2 y_2$, we arrive at the equations:

$$\begin{aligned} 0 &= y_1 u_1' + y_2 u_2' \\ f &= y_1' u_1' + y_2' u_2'. \end{aligned}$$

  - (Note that in order to arrive at these equations, we needed to start with an equation in *normalized form* $y'' + ay' + by = 0$, unlike in the previous example.)

- We can solve these equations explicitly by inverting the relevant 2×2-matrix. Recalling the Wronskian determinant $w = y_1 y_2' - y_2 y_1'$, the result is:

$$u_1'(x) = \frac{-y_2(x) f(x)}{w(x)} \qquad u_2'(x) = \frac{y_1(x) f(x)}{w(x)}.$$

- Integrating, we thus obtain

$$y_\mathrm{p}(x) = y_1(x) u_1(x) + y_2(x) u_2(x) = y_1(x) \int \frac{-y_2(x) f(x)}{w(x)} \, \mathrm{d}x + y_2(x) \int \frac{y_1(x) f(x)}{w(x)} \, \mathrm{d}x.$$

- The solution to the original normalized equation is then

$$y(x) = c_1 y_1(x) + c_2 y_2(x) + y_\mathrm{p}(x)$$

  with y_p as above and with c_1, c_2 *constants*.

  - It is nice to see that we can write out the solution explicitly like this, but it may be simpler just to remember the system of equations in u_1' and u_2' and start from there.

- If we fix $x_0 \in \mathbb{R}$, we can choose specific anti-derivatives in the expression for y_p above and obtain

$$y_\mathrm{p}(x) = \int_{x_0}^{x} \frac{y_1(t) y_2(x) - y_2(t) y_1(x)}{w(t)} f(t) \, \mathrm{d}t.$$

  - Moreover, we see that this particular $y_\mathrm{p}(x)$ is a solution (in fact, the unique solution, but we haven't proven this yet) to the initial value problem $Ly = f;\ y(x_0) = y'(x_0) = 0$.
  - The function $G(x,t) = \frac{y_1(t) y_2(x) - y_2(t) y_1(x)}{w(t)}$ is called the Green's function associated to the differential operator L.
  - Thus, to summarize, for any function f, integrating f against the Green's function $y(x) = \int_{x_0}^{x} G(x,t) f(t) \, \mathrm{d}t$ produces a solution to the IVP $Ly = f;\ y(x_0) = y'(x_0) = 0$.

Example 11.3.8-9

- We consider $Ly = x^2y'' - 2xy' + 2y = x^3$.
  - This is the same L we studied before, but now we put it in normalized form $y'' - \frac{2}{x}y' + \frac{2}{x^2}y = x$; since now we have x in a denominator, let us seek solutions $y\colon (0,\infty) \to \mathbb{R}$ or $(0,\infty) \to \mathbb{C}$.
  - More generally, let us solve $y'' - \frac{2}{x}y' + \frac{2}{x^2}y = f$ for an arbitrary f.
- Referring to example 11.3.7, the solutions to the associated homogeneous equation are x and x^2.
  - (In fact, this example is actually a bit silly, since, as we mentioned there, we will never need the second function x^2. Thus, one could just start with the *Ansatz* $y(x) = xu(x)$ and plug this into the equation as we did before. Nonetheless, we will use this example to demonstrate the general approach using both u_1 and u_2.)
- We thus seek a particular solution of the form $y_\mathrm{p}(x) = xu_1(x) + x^2u_2(x)$ (and will then have the general solution $y = y_\mathrm{p}(x) + c_1x + c_2x^2$).
  - Let us do this in three (more or less equivalent) ways, just to summarize the above discussion.
- First way: we set up the equations for u_1' and u_2':

$$\begin{aligned} xu_1' + x^2u_2' &= 0 \\ u_1' + 2xu_2' &= f \end{aligned}$$

  - Subtracting the first equation from x times the second, we obtain $x^2u_2' = fx$, and inserting this into the first equation gives $xu_1' = -fx$.
  - We thus have $u_1(x) = -\int f(x)\,\mathrm{d}x$ and $u_2(x) = \int \frac{f(x)}{x}\,\mathrm{d}x$.
  - We conclude that we have a particular solution

$$y_\mathrm{p}(x) = -x\int f(x)\,\mathrm{d}x + x^2\int \frac{f(x)}{x}\,\mathrm{d}x$$

  - For example when $f(x) = x$ as above, we obtain

$$y_\mathrm{p}(x) = -\frac{1}{2}x^3 + x^3 = \frac{1}{2}x^3.$$

- Second way: we compute the Wronskian

$$w(x) = y_1y_2' - y_2y_1' = 2x^2 - x^2 = x^2.$$

  - We now use the explicit formulas above for u_1 and u_2:

$$u_1(x) = \int \frac{-y_2(x)f(x)}{w(x)}\,\mathrm{d}x = \int \frac{-x}{x^2}f(x)\,\mathrm{d}x = -\int f(x)\,\mathrm{d}x$$

and

$$u_2(x) = \int \frac{y_1(x)f(x)}{w(x)}\,\mathrm{d}x = \int \frac{x}{x^2}f(x)\,\mathrm{d}x = \int \frac{f(x)}{x}\,\mathrm{d}x,$$

which of course gives the same answers.

- Third way: we compute the Green's function of L

$$G(x,t) = \frac{y_1(t)y_2(x) - y_2(t)y_1(x)}{w(t)} = \frac{tx^2 - t^2x}{t^2} = \frac{x^2}{t} - x.$$

  - We then obtain a particular solution y_p with initial conditions $y_\mathrm{p}(0) = y_\mathrm{p}'(0) = 0$ by integrating against the Green's function:

$$y_\mathrm{p}(x) = \int_0^x G(x,t)f(t)\,\mathrm{d}t = \int_0^x \Big(\frac{x^2}{t} - x\Big)f(t)\,\mathrm{d}t,$$

    which is again the same answer we got before (though written slightly differently).

Higher-order equation

- The method of variation of parameters also works for higher-order linear equations. We comment on it very briefly here.
- Given an n-th order linear equation $Ly = f$ (with possibly non-constant coefficients), we first find n solutions $y_1, \dots, y_n$ to the homogeneous equation.
- We then make the ansatz $y = u_1y_1 + \cdots + u_ny_n$ for some undetermined functions $u_1(x), \dots, u_n(x)$.
- Plugging this into $Ly = f$ and making a similar (but more complicated) computation to before, we end up with a sufficient condition for y to be a solution, namely that the functions $u_1', \dots, u_n'$ satisfy the linear system of equations (with non-constant coefficients)

$$A(x)\mathbf{U}(x) = \mathbf{b}(x),$$

  where $\mathbf{U}$ is the column vector $(u_1', \dots, u_n')$, $\mathbf{b}$ is the column vector $(0, \dots, 0, f)$, and A is the matrix with entries $A_{ij} = y_j^{(i)}$.
- When $n = 2$, this specializes to the system that we saw before.
- The determinant $w(x)$ of $A(x)$ is again called the *Wronskian* of the functions $y_1, \dots, y_n$, and we again have that it is non-zero for all x if the y_i are independent solutions to the homogeneous equation.
- Hence, we can solve invert A and solve the system for the u_i', and hence again we can conclude that this Ansatz will produce a solution.

18 Apr 7, 11.5: Laplace transforms

18.1 11.5: Laplace transforms

- Given a function $f\colon [0,\infty) \to \mathbb{R}$, its Laplace transform is the function $\mathcal{L}[f]$ defined by

$$\mathcal{L}[f](s) = \int_0^\infty e^{-st} f(t)\,\mathrm{d}t$$

  assuming this integral converges – otherwise the Laplace transform of f does not exist.

  - In general, the function $\mathcal{L}[f](s)$ is not defined for all values of s but only for s sufficiently large (as we'll soon see).
  - We recall that, in general, an improper integral $\int_a^\infty F(x)\,\mathrm{d}x$ is defined as the limit (if it exists) $\lim_{T\to\infty} \int_a^T F(x)\,\mathrm{d}x$.
  - Note that we can also define the Laplace transform for a function $f\colon \mathbb{R} \to \mathbb{R}$, but from the definition, it is clear that it only depends on the restriction of f to $[0,\infty)$.
  - In applications, one often says that the Laplace transform transforms a function "on the time domain" to one "on the frequency domain". This is because, if t has units of time, then s has units of inverse time, i.e., frequency, in order for the argument $-st$ of the exponential function to be dimensionless.

- Important example: for $a \in \mathbb{R}$, we have

$$\mathcal{L}[e^{at}](s) = \int_0^\infty e^{-st}e^{at}\,\mathrm{d}t = \lim_{T\to\infty}\int_0^T e^{(s-a)t}\,\mathrm{d}t = \lim_{T\to\infty}\left[-\frac{e^{(s-a)t}}{s-a}\right]_{t=0}^T = \frac{1}{s-a}.$$

  - We see here that the last limit converges only when $s > a$, and hence $\mathcal{L}[e^{at}]$ is defined only on (s,∞).

Key properties of the Laplace transform

- There are a couple of important properties that make the Laplace transform useful for solving differential equations.
- The first is that $\mathcal{L}$ behaves well with respect to derivatives. Namely, we have the formula:

$$\mathcal{L}[f'](s) = -f(0) + s\mathcal{L}[f](s)$$

  - For this formula to hold, f has to satisfy two conditions:
    - (i) The Laplace transforms $\mathcal{L}[f]$ and $\mathcal{L}[f']$ must exist (of course)
    - (ii) We must have $\lim_{t\to\infty} e^{-st}f(t) = 0$ for all (sufficiently large) s. (This is saying that f "does not grow too fast".)
  - The proof is a straightforward computation using integration by parts:

$$\begin{aligned}\mathcal{L}[f'](s) &= \lim_{T\to\infty}\int_0^T e^{-st}f'(t)\,\mathrm{d}t \\ &= \lim_{T\to\infty}\left[e^{-st}f(t)\right]_{t=0}^T + s\int_0^T e^{-st}f(t)\,\mathrm{d}t \\ &= -f(0) + s\mathcal{L}[f](s),\end{aligned}$$

    where in the last line we used assumption (ii).

- The second is that $\mathcal{L}$ is *linear*: $\mathcal{L}[c_1 f_1 + c_2 f_2] = c_1\mathcal{L}[f_1] + c_2\mathcal{L}[f_2]$ – this is immediate from the definition.

  - (Though we note that it is a bit awkward to formulate $\mathcal{L}$ as a linear map between two specific vector spaces, since its codomain is something like "the set of functions defined on (a, ∞) or $[a, \infty)$ for some $a \in [-\infty, \infty)$".)

- The third, very important property, is called **Lerch's theorem**: it says that if $\mathcal{L}[f] = \mathcal{L}[g]$, then $f = g$.

  - It's important to note that, if f and g are functions defined on all $\mathbb{R}$, then the conclusion of this theorem can nonetheless only be that f and g agree when restricted to $[0, \infty)$, since we can arbitrarily modify f and g on $(-\infty, 0)$ without changing their Laplace transforms.

  - This theorem more or less amounts to the existence of an *inverse Laplace transform* $\mathcal{L}^{-1}$, with $\mathcal{L}^{-1}[\mathcal{L}[f]] = f$. In the above example, for instance, we would have $\mathcal{L}^{-1}[\frac{1}{s-a}](t) = e^{at}$.

  - We will not prove Lerch's theorem. But we note that the nicest proof involves some auxiliary concepts, and we will comment more on them below.

Example 11.5.1

- Let's see how to put the above properties together to help us solve a differential equation; we start with a simple example – in fact, too simple to really show the usefulness of the methods, but it still illustrates the general idea.

- We seek a solution $y\colon \mathbb{R} \to \mathbb{R}$ to the IVP

$$y' + 2y = 0;\ y(0) = 3$$

- Applying $\mathcal{L}$ to both sides of the equation and using linearity and $\mathcal{L}[y'] = -y(0) + \mathcal{L}[y]$, we obtain

$$-3 + s\mathcal{L}[y](s) + 2\mathcal{L}[y](s) = 0.$$

- Hence, solving for $\mathcal{L}[y]$, we obtain $\mathcal{L}[y] = \frac{3}{s+2}$.

- We thus see that $\mathcal{L}[y] = \mathcal{L}[3e^{-2t}]$ and hence, by the existence of $\mathcal{L}^{-1}$, that $y(t) = 3e^{-2t}$.

- **Warning**: note that we can actually only conclude from this that $y(t) = 3e^{-2t}$ *for* $t \geq 0$!

  - Of course, we can easily check that this *is* in fact a solution for *all* t – though if we want to say that it's the *unique* solution, we can a priori only say this for $t \geq 0$.

  - However, it follows from the uniqueness theorem for first order ODEs that y is the unique solution on all of $\mathbb{R}$.

  - We will be able to draw a similar conclusion when applying the Laplace transform to higher-order equations using the uniqueness theorem for higher-order equations (which so far we have only stated in the case of linear equations with constant-coefficients).

Further remarks on the Laplace transform

- The Laplace transform is used heavily in electrical engineering, specifically control theory and signal processing.
- In those contexts, the restriction to functions $[0, \infty) \to \mathbb{R}$ often arises naturally (for example, when considering a signal $f(t)$ that only begins at time $t = 0$.
- It is possible, and sometimes necessary, to consider the Laplace transform $\mathcal{L}[f](s)$ $\int_0^\infty e^{-st} f(t)\,\mathrm{d}t$ as a function of a *complex* variable $s \in \mathbb{C}$. It follows from certain theorems of complex analysis that the values of $\mathcal{L}[f](s)$ with $s \in \mathbb{C}$ are completely determined by the restriction of $\mathcal{L}[f]$ to $s \in \mathbb{R}$.
- Another generalization which is important in applications is to consider $\mathcal{L}[f]$ not just for *functions* f but for *Schwartz distributions* such as the Dirac delta function δ.
  - This is a "function" with the property that $\int_{-\infty}^{\infty} \delta(x) f(x)\,\mathrm{d}x = f(0)$ for any (reasonable) function f, and in particular $\int_{-\infty}^{\infty} \delta(x)\,\mathrm{d}x = 1$.
  - Thus $\delta(x)$ is "completely concentrated" at the point $x = 0$ but has "total mass" 1.
  - This arises for example in signal processing in the analysis of *discrete*, rather than continuous, signals.
  - Similarly, for any $a \in \mathbb{R}$, there is a shifted delta function $\delta(x - a)$ with $\int_{-\infty}^{\infty} \delta(x - a) f(x)\,\mathrm{d}x$.
  - It follows for $a \geq 0$ that $\mathcal{L}[\delta(t - a)](s) = e^{-as}$.
- The Laplace transform is closely related to the *Fourier transform* $\mathcal{F}[f](s) = \int_{-\infty}^{\infty} e^{-2\pi i s t} f(t)\,\mathrm{d}t$, which is also used in a wide variety of applications, and is notably of central importance in quantum mechanics, and which has similar nice properties which make it useful in the study of differential equations.
- An explicit formula for the inverse Laplace transform $\mathcal{L}^{-1}[F]$ (when it exists) is given by

$$\mathcal{L}^{-1}[F](t) = e^{at} \int_{-\infty}^{\infty} F(a + 2\pi i s) e^{2\pi i t s}\,\mathrm{d}s.$$

  for an appropriate value of the constant $a \in \mathbb{R}$.
  - As you can see, this makes use of values of the Laplace transform $F = \mathcal{L}[f]$ as a function of a complex variable.
  - This is most easily proven using related properties of the Fourier transform.

Example 11.5.2

- The same approach as used above works for higher-order equations.
- The key fact is that

$$\mathcal{L}[y'](s) = -y'(0) + s\mathcal{L}[y'](s) = -y'(0) - sy(s) + s^2\mathcal{L}[y](s)$$

  and similarly (by induction)

$$\mathcal{L}[y^{(n)}](s) = s^n \mathcal{L}[y](s) - \sum_{k=0}^{n-1} s^{n-k-1} y^{(k)}(0).$$

– To use this, we need the same assumptions as above, which now say (i) all of the Laplace transforms $\mathcal{L}[y^{(k)}]$ exist for $k = 0, \dots, n$ and (ii) $\lim_{t\to\infty} e^{-st}y^{(k)}(t)$ for all (sufficiently large) s and all $k = 0, \dots, n-1$.

- Let's consider the IVP
$$y'' - y' - 2y = 3e^t;\ y(0) = 1;\ y'(0) = 0.$$
  – Setting $Y = \mathcal{L}[y]$ and applying $\mathcal{L}$ to both sides, we obtain
$$\begin{aligned}\frac{3}{s-1} &= (-y'(0) - sy(0) + s^2Y) - (-y(0) + sY) - 2Y \\ &= (s^2 - s - 2)Y - (s-1).\end{aligned}$$
  – Solving for Y, we have
$$Y = \frac{s^2 - 2s + 4}{(s-1)(s^2 - s - 2)}.$$
  – It remains to find the inverse Laplace transform of Y, for which it useful to first find a partial fraction decomposition: set
$$Y = \frac{A}{s-1} + \frac{B}{s+1} + \frac{C}{s-2}.$$
  – We find A by multiplying both sides by $s-1$ and setting $s = 1$, and we find B and C similarly.
  – This gives $A = -\frac{3}{2}$, $B = \frac{7}{6}$, $C = \frac{4}{3}$ and hence
$$Y(s) = \frac{-3/2}{s-1} + \frac{7/6}{s+1} + \frac{4/3}{s-2}.$$
  – Applying the inverse Fourier transform, we thus have
$$y(t) = -\frac{3}{2}e^t + \frac{7}{6}e^{-t} + \frac{4}{3}e^{2t}.$$
  – Again, **this holds a priori only for** $t \geq 0$, but by the uniqueness theorem for second order ODEs (which we actually have in this case, since this is a linear equation with constant coefficients), it follows that it holds for all t.

Tables of Laplace transforms

- We now have a feeling for the general method: apply the Laplace transform to both sides of a differential equation in y to turn it into an algebraic equation in $Y = \mathcal{L}[y]$, and then solve for Y.
- Then, after possibly rewriting Y as a linear combination of simpler functions (e.g., by partial fraction decomposition), find the inverse Laplace transform of Y (by finding it for each of the simpler pieces).
- This last step can be facilitated by consulting some big table of known Laplace transforms; one such table is provided in the textbook, in Table 111.1 on p. 545.
- There are also various other properties of the Laplace transform and inverse Laplace transform which makes them easier to compute.
  – Prominent among these (and discussed in section 11.6 in the book) is that (under certain assumptions) *the Laplace transform turns convolution into multiplication*:
$$\mathcal{L}[f * g](s) = \mathcal{L}[f](s) \cdot [g](s).$$

– Here, the convolution of two functions $f, g\colon [0,\infty) \to \mathbb{R}$ is defined by

$$(f * g)(t) = \int_0^t f(u)g(t-u)\,\mathrm{d}u.$$

19 Apr 9, 12.1: Vector fields

19.1 12.1A: Geometric interpretation

- We now begin our study of systems of differential equations.
- In general, such systems will have several unknowns, each of which is a function which together with its derivatives appears in the given equations.
- We will now change our notation and systematically use t rather than x for the variable with respect to which derivatives our taken.
  - This is so that we can use $x(t)$ and $y(t)$ for the unknowns in a given system. As is common when considering functions of time, we will also begin to use Newton's notation $\dot{x}(t)$ and $\dot{y}(t)$ for differentiation, rather than Lagrange's notation $x'(t)$ and $y'(t)$ as we have been doing. Thus, a typical first-order system of two equations in the unknowns $x(t)$ and $y(t)$ would look like
  $$\dot{x}(t) = F(t, x, y)$$
  $$\dot{y}(t) = G(t, x, y)$$
  - We write $\ddot{x}$ for the second derivative, but still write $x^{(n)}$ for higher derivatives.

Example 12.1.1

- Consider the system
$$\dot{x} = x$$
$$\dot{y} = 2y$$
- We seek a solution (x, y) with $x, y \colon \mathbb{R} \to \mathbb{R}$
- This is the simplest kind of system, since each unknown occurs in just one equation (and each equation contains a single unknown). such a system is called **uncoupled**.
- We thus solve it by solving each equation separately; so here, the general solution is (x, y) where $x(t) = c_1 e^t$ and $y(t) = c_2 e^{2t}$.

The geometric interpretation

- As usual, it is convenient to use vector notation when dealing with several variables.
- Thus, in the above example, we combine x and y into a function $\mathbf{x} \colon \mathbb{R} \to \mathbb{R}^2$ given by $\mathbf{x}(t) = (x(t), y(t))$. The system then becomes the single vector equation
$$\dot{\mathbf{x}}(t) = F(x(t), y(t)).$$
where $F \colon \mathbb{R}^2 \to \mathbb{R}^2$ is the function $F(x, y) = (x, 2y)$.
- To give a coupled example, the system
$$\dot{x} = x + y + t$$
$$\dot{y} = x - y - t$$
corresponds to the single equation
$$\dot{\mathbf{x}} = F(t, x, y)$$
where $F(t, x, y) = (x + y + t, x - y - t)$.

- Note that a solution to such an equation is a *plane curve*, i.e., a function $\mathbf{x}\colon \mathbb{R} \to \mathbb{R}^2$; for a system in n unknowns, it would be a curve in $\mathbb{R}^n$.
  - Such a curve in general is called a **trajectory** of the system.
  - As we saw, for the example considered above, the trajectories are exactly the curves of the form $(x, y) = (c_1 e^t, c_2 e^{2t})$.
  - If we fix an initial condition, say $(x(0), y(0)) = (1, 2)$, then this isolates a particular trajectory $(x, y) = (e^t, 2e^{2t})$.
  - Note that this trajectory is entirely inside the first quadrant, and as $t \to -\infty$, it approaches the origin, and it is in fact just the right (open) half of the parabola $y = 2x^2$.
- We can make a sketch of several trajectories with varying initial conditions c_1 and c_2 to get a feel for the general solution of the equation; such a picture is called a **phase portrait**.
  - (Note however, that the phase portrait is incomplete in an important sense: the trajectories are *parametrized* curves, having a definite *velocity* at each point, which is not captured in the phase portrait.)
  - In this case, there are three types of curves in the phase portrait: there are half-parabolas in each quadrant, all converging to 0 as $t \to -\infty$; there are the positive and negative x and y axes; and finally, when $c_1 = c_2 = 0$, there is simply the constant solution $(x(t), y(t)) = (0, 0)$. This last type of solution is called an **equilibrium solution**, and we say that the origin is an **equilibrium point** of this system.

General definition of a system of equations

- We can give a formal definition of a system of ordinary differential equation by repeating the definition of a single differential equation, simply replacing everything with vectors.
- Recall that we formally defined an ordinary differential equation of order m simply to be a function $F\colon \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}$ (or maybe $F\colon U \to \mathbb{R}$ for some $U \subset \mathbb{R}^{m+1}$), which we think of as representing the equation
$$x^{(m)}(t) = F(t, x(t), \dot{x}(t), \ldots, x^{(m-1)}(t)).$$
- Thus, in just the same way, we can formally define an m-th order system of differential equations in n variables to be given by an arbitrary function $F\colon \mathbb{R} \times (\mathbb{R}^n)^m \to \mathbb{R}^n$ (or more generally $F\colon U \to \mathbb{R}^n$ for some $U \subset \mathbb{R} \times (\mathbb{R}^n)^m$), which we think of as representing the equation
$$\mathbf{x}^{(m)}(t) = F(t, \mathbf{x}(t), \dot{\mathbf{x}}(t), \ldots, \mathbf{x}^{(m-1)}(t)).$$
- A solution to this differential equation on an interval $I \subset \mathbb{R}$ then is a n-times differentiable function $\mathbf{x}\colon I \to \mathbb{R}^n$ satisfying
$$\mathbf{x}^{(m)}(t) = F(t, \mathbf{x}(t), \dot{\mathbf{x}}(t), \ldots, \mathbf{x}^{(m-1)}(t))$$
for all $t \in I$ (where, in the more general situation involving a set U, we must also require that $(t, \mathbf{x}(t), \dot{\mathbf{x}}(t), \ldots, \mathbf{x}^{(m-1)}(t)) \in U$ for all $t \in I$).

- Writing $\mathbf{x}(t) = (x_1(t), \ldots, x_m(t))$ and similarly writing $F_1, \ldots, F_n$ for the components of F, this corresponds to n equations

$$\begin{aligned} x_1^{(m)}(t) &= F_1(t, \mathbf{x}(t), \dot{\mathbf{x}}(t), \ldots, \mathbf{x}^{(m-1)}(t)) \\ &\vdots \\ x_n^{(m)}(t) &= F_n(t, \mathbf{x}(t), \dot{\mathbf{x}}(t), \ldots, \mathbf{x}^{(m-1)}(t)) \end{aligned}$$

- Thus, for instance, if we consider a second-order system in three unknowns, we can write $\mathbf{x}(t) = (x(t), y(t), z(t))$ and

$$F(t, \mathbf{x}, \dot{\mathbf{x}}) = (F_1(t, x, y, z, \dot{x}, \dot{y}, \dot{z}), F_2(t, x, y, z, \dot{x}, \dot{y}, \dot{z}), F_3(t, x, y, z, \dot{x}, \dot{y}, \dot{z}))$$

and the equation $\ddot{\mathbf{x}} = F(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))$ then corresponds to the system

$$\begin{aligned} \ddot{x} &= F_1(t, x, y, z, \dot{x}, \dot{y}, \dot{z}) \\ \ddot{y} &= F_2(t, x, y, z, \dot{x}, \dot{y}, \dot{z}) \\ \ddot{z} &= F_3(t, x, y, z, \dot{x}, \dot{y}, \dot{z}) \end{aligned}$$

19.2 12.1B: Autonomous systems

- We now consider a first order equation $\dot{\mathbf{x}} = F(t, \mathbf{x})$.
- Recall that, geometrically, the vector $\dot{\mathbf{x}}(t)$ is the **velocity vector** of the curve $\mathbf{x}$, and is a vector tangent to curve $\mathbf{x}$ at the point $\mathbf{x}(t)$. (For this reason, the velocity vector is also called a **tangent vector**, though it's to be remembered that the velocity vector also contains the information of the *speed* (which could be 0!), and not just the direction).
- Thus, we see that the equation is simply specifying what the velocity vector of a trajectory should be at each point.
- In other words, the function $F\colon \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ defining a first-order equation is simply a *vector field*, and a solution is simply a parametrized curve whose velocity vector is always the one specified by the vector field (such a curve is sometimes called an **integral curve** or **flow line** of the vector field). It should be noted that if $F(t, \mathbf{x})$ really depends on t, that this is a *time-varying* vector field.
- A (possibly higher order) system of ODEs which does not depend on time is called autonomous.

Example 12.1.2

- An example of an autonomous system is $(\dot{x}, \dot{y}) = (-y, x)$.

$$\begin{aligned} \dot{x} &= -y \\ \dot{y} &= x \end{aligned}$$

- The resulting vector field is depicted in Figure 12.4 (b).
- It strongly suggests that the trajectories of the system are circles around the origin.
- In fact, a naive approach to solving this system suggests itself: substitute one equation in the other.
  - This gives $\ddot{x} = -x$, which we know has (among others) solutions $x = r\cos t$.
  - Substituting this in the second equation gives $\dot{y} = x = r\cos t$, which has a solution $y = r\sin t$.
  - We see the circles $(x, y) = (r\cos t, r\sin t)$ are indeed trajectories of the system.
- We will see soon more systematic methods for solving this system.

Another example

- As we noted when discussing slope fields, they are related to, but not the same as vector fields, and one should make the distinction clear to oneself.
- However, the two concepts can be related as follows.
- Given a first-order equation $\dot{y} = f(t, y)$, one can define the system

$$\begin{aligned} \dot{x} &= 1 \\ \dot{y} &= f(x, y). \end{aligned}$$

- In other words, this is the vector field $F(x, y) = (1, f(x, y))$.
- Because all of the vectors have x-coordinate 1, they are determined by their slope $f(x, y)$, and so this essentially amounts to a slope field.
- Moreover, the trajectories of this system are precisely the graphs of the solutions the original equation.

Example 12.1.3

- Here is an example of a time-dependent vector field (hence of a non-autonomous system):

$$(\dot{x}, \dot{y}) = \big((1-t)x - ty, tx + (1-t)y\big).$$

- At $t = 1$, this is precisely the vector field we considered before.
  - But for instance at $t = 0$, it is radially outward.
  - There are some sketches in Figure 12.5.
- It is much less straightforward to solve this system.

Example 12.1.4

- Suppose that in $(\dot{x}, \dot{y}) = (F(t, x, y), G(t, x, y))$, the functions F and G are independent of t, or more generally that $R := F/G$ is (non-zero and) independent of t.
- Suppose moreover that we have a trajectory $(x(t), y(t))$ such that y is a function of x, so that $y(t) = f(x(t))$ (by the implicit function theorem, this will happen on some interval $(t_0 - \varepsilon, t_0 + \varepsilon)$ for any t_0 with $\dot{x}(t_0) \neq 0$).
- Then by the chain rule, we have

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\mathrm{d}y/\,\mathrm{d}t}{\mathrm{d}x/\,\mathrm{d}t} = \frac{F}{G} = R(x, y).$$

- This can give us some information about the (unparametrized) trajectory curves.
- For instance, if $(F, G) = (ty, -tx)$, we obtain $R = -x/y$, and hence

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{-x}{y}.$$

- By separation of variables, we find that $x^2 + y^2 = c$ for some $c \in \mathbb{R}$.
- So we see that any part of a trajectory away from the x-axis and with non-zero velocity in the x direction must lie on some fixed circle.

20 Apr 14, 12.1C: higher-order systems

20.1 12.1C: Second-Order equations

- We now introduce an important trick which shows that – surprisingly – a higher-order system of ODEs can always be reduced to a first-order system.
- The price to pay is that the number of equations increases; in particular, if you start with a single n-th order equation, you end up with a system of n linear equations.
- We first consider the second-order case, so we have an equation

$$\ddot{y} = f(t, y, \dot{y}).$$

  - Now introduce a new function $z(t) = \dot{y}(t)$. We see that the pair (y, z) satisfies the first-order system

$$\begin{aligned} \dot{y} &= z \\ \dot{z} &= f(t, y, z). \end{aligned}$$

  - Conversely, given any solution (y, z) to this system, we y will be a solution to the original equation; hence we see that the original equation and the new system of equations are equivalent.

- We can also carry over initial conditions: the IVP

$$\ddot{y} = f(t, y, \dot{y});\ y(t_0) = y_0,\ \dot{y}(t_0) = z_0$$

  corresponds to the system

$$\begin{aligned} \dot{y} &= z & y(t_0) &= y_0 \\ \dot{z} &= f(t, y, z) & z(t_0) &= z_0. \end{aligned}$$

- As an example, consider the harmonic oscillator $\ddot{x} + x = 0$.

  - This turns into the system

$$\begin{aligned} \dot{x} &= y \\ \dot{y} &= -x \end{aligned}$$

    we saw above.

  - As we saw above, the solutions to this system are circular trajectories. But now we view them with a new perspective: they are simultaneously plotting the position and velocity of a harmonic oscillator.

- It's clear how to generalize this to higher orders: the equation

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

  is transformed into the system

$$\begin{aligned} \dot{x}_0 &= x_1 \\ \dot{x}_1 &= x_2 \\ &\vdots \\ \dot{x}_{n-2} &= x_{n-1} \\ \dot{x}_{n-1} &= f(t, x_0, \ldots, x_{n-1}) \end{aligned}$$

  - Then y is a solution to the original equation if and only if $y, \dot{y}, \ldots, y^{(n-1)}$ is a solution to the new system.
  - It is also clear how to transform an initial value problem, and also how to transform a higher-order system with possibly many equations.

- It is very good to know theoretically that every system is equivalent to a first-order system; though when trying to solve or analyze a particular higher-order system, it may or may not be helpful to convert it into a first-order one.

A note on "characteristic equations"

- A couple of people in class have asked what if any connection there is between the "characteristic polynomial" of a linear differential operator, and the characteristic polynomial of a matrix from linear algebra
- We will discuss the characteristic polynomial soon: it is a degree n polynomial $p_A(x)$ associated to any square matrix $A \in K^{n \times n}$.
- We can see one connection by applying the above procedure to turn a single n-th order linear ODE into a system of n first-order linear ODEs.
- Namely, any such system (say, with constant coefficients) has the form $\dot{\mathbf{x}} = A\mathbf{x} + \mathbf{b}$, with $A \in K^{n \times n}$ and $\mathbf{b} \in K^n$, and, we can thus consider the characteristic polynomial A, which is a
  - Soon, we will have much more to say about studying a system of linear ODEs in terms of the matrix A and its characteristic polynomial.
- Now if you apply this construction to a linear ODE $\dot{x}^{(n)} + a_{n-1}x^{(n-1)} + \cdots + a_1\dot{x} + a_0 x = 0$ with characteristic polynomial $p(r)$, you end up with the matrix

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

- This is precisely the so-called **companion matrix** of the polynomial p, and it is easy to see check by induction (once one has the definition of the characteristic polynomial) that it's character polynomial is p.
- Thus, we see that the characteristic polynomial of a higher-order linear equation is just the characteristic polynomial of the matrix representing the corresponding first-order system of equations.

Existence and uniqueness theorem

- We now state the general existence and uniqueness theorem for higher-order equations and systems of equations.
  - Or rather, we will only state a theorem about first-order systems of equations, but by the above trick, this immediate implies a corresponding theorem about higher-order equations and systems!

- The theorem statement is a direct generalization of the theorem we saw for a single first-order equation:
- **Theorem**: Suppose $F\colon U \to \mathbb{R}^n$ is continuous, where $U = I_1 \times V \subset \mathbb{R} \times \mathbb{R}^n$ for some open interval I_1 and some open set $V \subset \mathbb{R}^n$. Suppose that the derivative $F_{\mathbf{x}}(t, \mathbf{x})$ with respect to $\mathbf{x}$ (i.e., the Jacobian matrix $(\frac{\partial F_i}{\partial x_j})(t, \mathbf{x})$) exists and is continuous.
  - Then for any $(t_0, \mathbf{x}_0) \in U$, the IVP $\dot{\mathbf{x}} = F(t, \mathbf{x});\ \mathbf{x}(t_0) = \mathbf{x}_0$ has a solution defined on some interval $I \subset I_1$, and this solution is unique in the sense that for any two solutions $\mathbf{x}_1\colon I \to \mathbb{R}$ and $\mathbf{x}_2\colon I' \to \mathbb{R}$, we have $\mathbf{x}_1(t) = \mathbf{x}_2(t)$ for $t \in I \cap I'$.
  - If moreover $V = \mathbb{R}$ and there is some $B > 0$ such that all of the entries of $F_{\mathbf{x}}(t, \mathbf{x})$ are bounded by B for all $(t, \mathbf{x}) \in U$, then the solution will exist on the entire interval I_1.
- Not only is the statement the same as in the one-dimensional version of the theorem, the proof (which we only sketched) is the same as well.
  - That is, one again converts the differential equation into an integral equation $\mathbf{x}(t) = \mathbf{x}_0 + \int_{t_0}^{t} F(t, \mathbf{x}(t))\,\mathrm{d}t$.
  - Then one defines a solution again as the limit of the Picard iteration of this integral, and by the same arguments (which are non-trivial and which I didn't say anything about), one shows that this limit exists and is a solution.
- **Warning**: one must be careful to interpret the uniqueness in the case of *non-autonomous* system.
  - For an autonomous system, the uniqueness theorem essentially says that, for a given vector field (satisfying the assumptions of the theorem), any two trajectories passing through the same point must agree.
  - In particular, if a trajectory $\mathbf{x}(t)$ satisfies $\mathbf{x}(t_0) = \mathbf{x}(t_1)$ for some $t_0 < t_1$, then we must have $\mathbf{x}(t_0 + t) = \mathbf{x}(t_1 + t)$, hence $\mathbf{x}(t) = \mathbf{x}(t + t_1 - t_0)$ for all t, i.e., x is a *periodic trajectory* with period $t_1 - t_0$.
  - However, in a non-autonomous system, the uniqueness theorem only says two trajectories passing through the same point *at the same time* must agree.
  - In particular, a trajectory can cross itself at different times without being periodic.

20.2 A bit of 13.1 before turning to 3.6: eigenvalues and eigenvectors

- We now turn to the study of first-order *homogeneous* constant-coefficient systems of ODEs.
- These are precisely systems of the form
$$\dot{\mathbf{x}} = A\mathbf{x}$$
for some $A \in K^{n \times n}$.
- As in the one-dimensional situation, the study of these will be important also for the non-homogeneous and non-constant coefficients cases.
- Now, in the 1-dimensional case, this equation is very simple: it is just
$$\dot{x} = ax$$
with some $a \in \mathbb{R}$, with general solution $x = ue^{at}$ for some constant $u \in \mathbb{R}$.

- Later, we will see that, amazingly, this generalizes quite directly: there is a notion of *matrix exponential* e^A, and we will find solutions of the form $e^{At}\mathbf{u}$.
- For now, inspired by the 1-dimensional case, let us simply seek solutions of the form $x(t) = e^{\lambda t}\mathbf{u}$ for some $\lambda \in K$ and some non-zero $\mathbf{u} \in K^n$ (by which we mean $K^{n\times 1}$ – we will be using this shorthand frequently in what follows).
- We then have $\dot{x} = \lambda e^{\lambda t}\mathbf{u}$.
- Thus, this is a solution if and only if $A(e^{\lambda t}\mathbf{u}) = \lambda e^{\lambda t}\mathbf{u}$ which, because $\mathbf{u}$ is non-zero, is equivalent to

$$A\mathbf{u} = \lambda\mathbf{u}.$$

- **Definition**: given $A \in K^{n\times n}$, a non-zero vector $\mathbf{u} \in K^n$ is called an eigenvector of A with eigenvalue λ if $A\mathbf{u} = \lambda\mathbf{u}$. We say that $\lambda \in K$ is an eigenvalue of A if A has some eigenvector with eigenvalue λ.
  - (The prefix "eigen" in German means "proper" or "characteristic".)
  - Geometrically, this means that the linear transformation A, though it may be quite complicated in general, acts on vectors in the direction of $\mathbf{u}$ simply by scaling it by λ. (Note that if $\mathbf{u}$ is an eigenvector with eigenvalue λ, so is any non-zero scalar multiple of $\mathbf{u}$.)
  - The reason we exclude the zero vector $\mathbf{u} = \mathbf{0}$ is that it satisfies $A\mathbf{u} = \lambda\mathbf{u}$ for *every* λ. Hence, if we allowed it, every number would be an eigenvalue of A.
  - Note that if A is a *real* matrix, it might still have complex eigenvalues, in the sense that it has these eigenvalues when regarded as a complex matrix. An example is $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$ which has eigenvectors $\begin{bmatrix} 1 \\ \pm i \end{bmatrix}$ with eigenvalues $\mp i$.
- Thus, we see that $x = e^{\lambda t}\mathbf{u}$ is a solution if and only if $\mathbf{u}$ is an eigenvector of A with eigenvalue λ.
- We thus next turn to the general study of eigenvectors and eigenvalues.

21 Apr 16, 3.6: Eigenvalues and eigenvectors

21.1 3.6A: Definition and examples, and 3.6B: Bases of eigenvectors

- Above, we defined the eigenvectors and eigenvalues of a *matrix* $A \in K^{n\times n}$.
- More generally, suppose $L \in \mathcal{L}(V)$ is any operator on a K-vector space V.
  - We say that a non-zero vector $\mathbf{v} \in V$ is an eigenvector of L with eigenvalue λ (and that λ is an eigenvalue of L) if $L\mathbf{v} = \lambda\mathbf{v}$.
  - If $V \in K^n$, then we can consider L as a matrix in $K^{n\times n}$, and this definition then agrees with the previous one.
  - More generally, if V is finite-dimensional, and we consider the matrix $[L]_{\mathcal{B}}$ representing L with respect to any ordered basis $\mathcal{B}$, and writing $[\mathbf{v}]_{\mathcal{B}}$ for the coordinate vector of $\mathbf{v} \in V$ with respect to $\mathcal{B}$, we have $[L\mathbf{v}]_{\mathcal{B}} = [L]_{\mathcal{B}}[\mathbf{v}]_{\mathcal{B}}$ and hence $L\mathbf{v} = \lambda\mathbf{v}$ if and only if $[L]_{\mathcal{B}}[\mathbf{v}]_{\mathcal{B}} = \lambda[\mathbf{v}]_{\mathcal{B}}$ (since $[\lambda\mathbf{v}]_{\mathcal{B}} = \lambda[\mathbf{v}]_{\mathcal{B}}$).
  - It follows that $\mathbf{v}$ is an eigenvector of V with eigenvalue λ if and only if its coordinate vector $[\mathbf{v}]_{\mathcal{B}}$ is an eigenvector of $[L]_{\mathcal{B}}$ with respect to λ, and in particular that λ is an eigenvalue of L if and only if it is an eigenvalue of $[L]_{\mathcal{B}}$.

Example 3.6.4

- The differentiation operator $D \in \mathcal{L}(\mathcal{C}^\infty(\mathbb{R}))$ has eigenvectors e^{rx} with eigenvalue r.

Bases of eigenvectors

- Let $L \in \mathcal{L}(V)$ be a linear operator, and suppose $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are eigenvectors with eigenvalues $\lambda_1, \ldots, \lambda_k$, so that $L(\mathbf{v}_j) = \lambda_j\mathbf{v}_j$ for all j.
  - Then it is easy to compute the action of L on any linear combination of the $\mathbf{v}_j$'s.
  - Indeed, if $\mathbf{v} = c_1\mathbf{v}_1 + \cdots c_k\mathbf{v}_k$, we have by linearity:

$$L(\mathbf{v}) = c_1L\mathbf{v}_1 + \cdots c_kL\mathbf{v}_k = c_1\lambda_1\mathbf{v}_1 + \cdots + c_k\lambda_k\mathbf{v}_k.$$

- In particular, if the $\mathbf{v}_j$'s form a *basis* of V, then we can express any vector $\mathbf{v}$ as a linear combination of the $\mathbf{v}_j$'s and thence easily compute $L\mathbf{v}$.
- This is most neatly expressed by saying that if $\mathcal{B} = (\mathbf{v}_1, \ldots, \mathbf{v}_n)$ is an ordered basis of eigenvectors with eigenvalues $\lambda_1, \ldots, \lambda_n$, then the *matrix* $[L]_{\mathcal{B}}$ of L with respect to $\mathcal{B}$ is *diagonal*:

$$[L]_{\mathcal{B}} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}.$$

  - Indeed, this follows immediately since the j-th column of L is

$$[L]_{\mathcal{B}}\mathbf{e}_j = [L]_{\mathcal{B}}[\mathbf{v}_j]_{\mathcal{B}} = [L\mathbf{v}_j]_{\mathcal{B}} = \lambda_j[\mathbf{v}_j]_{\mathcal{B}} = \lambda_j\mathbf{e}_j.$$

  - We express this by saying that the operator L is diagonalizable.

- Conversely, if $[L]_{\mathcal{B}}$ is diagonal, then $\mathbf{e}_j$ is an eigenvector with eigenvalue λ_j for each j, hence $\mathbf{v}_j$ is as well, and so $\mathcal{B}$ is a basis of eigenvectors with eigenvalues $\lambda_1, \ldots, \lambda_n$.
- Thus, diagonalizability of L is exactly the condition that V have a basis of eigenvectors of L; this is **Theorem 3.6.5**.

- Geometrically, the diagonalizability of L means that there exists a coordinate system of V such that L simply scales the vectors on each coordinate axis (possibly by different amounts).

Example 3.6.1 and 3.6.2

- Consider the matrix $A = \begin{bmatrix} 1 & 1 \\ 4 & 1 \end{bmatrix}$.
- This has a basis of eigenvectors $(1, 2)$ and $(1, -2)$ with eigenvalues 3 and -1, respectively.
- In Figure 3.10, there is a depiction of the action of A on vectors in the plane.

6C: Change of basis matrices

- Given a matrix $A \in K^{n \times n}$ such that K^n has an ordered basis $\mathcal{B}$ of eigenvectors of A, we have the associated diagonal matrix $D = [A]_{\mathcal{B}}$.
  - We now ask: what is the relationship between A and D?
  - The answer is that they are *conjugate*; in general, two square matrices X and Y are conjugate if $Y = UXU^{-1}$ for some invertible matrix U.

- To see this, we consider a related question, given two ordered bases $\mathcal{B}$ and $\mathcal{B}'$ of a finite-dimensional vector space V, and a linear operator $L \in \mathcal{L}(V)$, what is the relationship between the matrices $[L]_{\mathcal{B}}$ and $[L]_{\mathcal{B}'}$?
  - (A similar question could be asked – and would receive a similar answer – about an arbitrary linear map $f\colon V \to W$ between finite-dimensional vector spaces, and given bases $\mathcal{B}, \mathcal{B}'$ of V and $\mathcal{C}, \mathcal{C}'$ of W.)

- Using the relationship between linear maps and matrix multiplication, we proceed by writing $L = \mathrm{id}_V \circ L \circ \mathrm{id}_V$, and concluding that $[L]_{\mathcal{B}'} = [\mathrm{id}_V]_{\mathcal{B}\mathcal{B}'}[L]_{\mathcal{B}}[\mathrm{id}_V]_{\mathcal{B}'\mathcal{B}}$.
  - Note that since $[\mathrm{id}_V]_{\mathcal{B}\mathcal{B}'}[\mathrm{id}_V]_{\mathcal{B}'\mathcal{B}} = [\mathrm{id}_V]_{\mathcal{B}} = \mathrm{I}$, it follows that $[\mathrm{id}_V]_{\mathcal{B}'\mathcal{B}} = [\mathrm{id}_V]^{-1}_{\mathcal{B}\mathcal{B}'}$, and hence that $[L]_{\mathcal{B}'}$ is a conjugate of $\mathcal{B}$.
  - Here, the matrix $[\mathrm{id}_V]_{\mathcal{B}\mathcal{B}'}$ is called a change-of-bases matrix. Its i-th column is the coordinate vector with respect to $\mathcal{B}'$ of the i-th basis vector in $\mathcal{B}$.

- Returning to the case of a matrix A, we have that $A = [A]_{\mathcal{E}}$, where $\mathcal{E} = (\mathbf{e}_1, \ldots, \mathbf{e}_n)$ is the standard basis.
  - It follows that for any basis $\mathcal{B}$, we have $[A]_{\mathcal{B}} = [\mathrm{id}]^{-1}_{\mathcal{B}\mathcal{E}} A [\mathrm{id}]_{\mathcal{B}\mathcal{E}}$. (This, together with the next statement, is **Theorem 3.6.8**.
  - Here, the matrix $[\mathrm{id}]_{\mathcal{B}\mathcal{E}}$ is particular simple: its columns are simply the elements of the basis $\mathcal{B}$ (each of which is an element of K^n, which we consider as a column vector).

– Conversely, we see that for *any* matrix U, we have that $U^{-1}AU$ is equal to $[A]_{\mathcal{B}}$ for some basis $\mathcal{B}$, namely the one whose elements are the columns of U.

- In conclusion, to say that a matrix A is diagonalizable is to say that $A = UDU^{-1}$ for some diagonal matrix D and some invertible matrix U (whose columns are then a basis of eigenvectors of A).

Example 3.6.14

- Considering the above example $A = \begin{bmatrix} 1 & 1 \\ 4 & 1 \end{bmatrix}$ with eigenbasis $(1, 2)$ and $(1, -2)$ with eigenvalues 3 and -1, we conclude that
$$A = U \begin{bmatrix} 3 & 0 \\ 0 & -1 \end{bmatrix} U^{-1}$$
where $U = \begin{bmatrix} 1 & 1 \\ 2 & -2 \end{bmatrix}$ and hence $U^{-1} = \frac{1}{-4} \begin{bmatrix} -2 & -1 \\ -2 & 1 \end{bmatrix} = \begin{bmatrix} 1/2 & 1/4 \\ 1/2 & -1/4 \end{bmatrix}$.

Finding eigenvalues, and review of determinants

- Let us see how to find the eigenvalues of a matrix A.
- The equation $A\mathbf{v} = \lambda\mathbf{v}$ expressing that $\mathbf{v} \neq \mathbf{0}$ is an eigenvector of A with eigenvalue λ can be rewritten as $(\lambda \mathrm{I} - A)\mathbf{v} = \mathbf{0}$, i.e., $\mathbf{v} \in \ker(\lambda \mathrm{I} - A)$.
- Thus, we see that λ **is an eigenvalue of** A **if and only** $\lambda \mathrm{I} - A$ **has a non-trivial kernel** or, what amounts to the same thing, if and only if $\lambda I - A$ is **not invertible**.
- Now, we recall that a criterion for invertibility of a matrix is the non-vanishing of its determinant. Let us recall the definition of the determinant.
- The nicest way to introduce the determinant is by first describing a simple property which determines it uniquely.

  – Given K-vector spaces $V_1, \ldots, V_n, W$, we call a map $F\colon V_1 \times \cdots \times V_n \to W$ <u>multilinear</u> if it is linear in each argument, when all of the other arguments are held fixed.

  – That is,
  $$F(a\mathbf{u}_1 + b\mathbf{w}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n) = aF(\mathbf{u}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n) + bF(\mathbf{w}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$$
  and
  $$F(\mathbf{v}_1, a\mathbf{u}_2 + b\mathbf{w}_2, \ldots, \mathbf{v}_n) = aF(\mathbf{v}_1, \mathbf{u}_2, \ldots, \mathbf{v}_n) + bF(\mathbf{v}_1, \mathbf{w}_2, \ldots, \mathbf{v}_n)$$
  and so on.

  – Next, in the case when all the V_i are the same vector space V, so that $F\colon V^n \to W$, we say that F is <u>alternating</u> if it changes sign whenever two of its arguments are swapped.

  – That is,
  $$F(\mathbf{v}_1, \ldots, \mathbf{v}_n) = -F(\mathbf{v}_1, \ldots, v_{i-1}, \mathbf{v}_j, \mathbf{v}_{i+1}, \ldots, \mathbf{v}_{j-1}, \mathbf{v}_i, \mathbf{v}_{j+1}, \ldots, \mathbf{v}_n).$$

  – Now any $(n \times n)$-matrix can be considered as a sequence of n column vectors. Hence any function $F\colon K^{n\times n} \to K$ can be regarded as a function $(K^n)^n \to K$, and thus we can talk about F being multilinear or alternating.

– The determinant $\det\colon K^{n\times n}$ is then **the unique alternating multilinear function satisfying** $\det \mathrm{I} = 1$.

- To see why there can be at most one such function det, let A be a matrix with columns $\mathbf{v}_1, \ldots, \mathbf{v}_n$.
  - We then have $\mathbf{v}_j = \sum_{i=1}^n A_{ij}\mathbf{e}_j$ for each j.
  - Using the multilinearity of A, it follows that

$$\det A = \det(\mathbf{v}_1, \ldots, v_n) = \sum_{i_1,\ldots,i_n=1}^{n} A_{i_1,1}\cdots A_{i_n,n}\det(\mathbf{e}_{i_1}, \ldots, \mathbf{e}_{i_n}).$$

  - Now, it follows immediately from the alternating property of det that $\det B = 0$ whenever a matrix B has two identical columns.
  - Hence, in the above sum, all the terms vanish except those in which the i_j are all distinct, and hence in which $(i_1, \ldots, i_n) = (\sigma(1), \ldots, \sigma(n))$ for some *bijection* $\sigma\colon \{1, \ldots, n\} \to \{1, \ldots, n\}$. Such a bijection is also known as a permutation, and the set of permutations is denoted Σ_n.
  - The sign $\operatorname{sgn}(\sigma)$ of a permutation $\sigma \in \Sigma_n$ is given by $\operatorname{sgn}(\sigma) = (-1)^N$, where N is the number of pairs of elements of $\{1, \ldots, n\}$ that get swapped by σ: $N = \#\{(i, j) \mid i < j \wedge \sigma(i) > \sigma(j)\}$.
  - Now, it is easy to see that any permutation is a composite of transpositions, meaning permutations which exchange two elements and leave all the others in place. One can then check by induction that the sign of a permutation $\sigma \in \Sigma_n$ is exactly the number of transpositions in any representation of σ as a composite of transpositions; i.e., if $\sigma = \tau_N \circ \cdots \circ \tau_1$ with each τ_i a transposition, then $\operatorname{sgn}(\sigma) = N$.
  - It follows that any alternating function F satisfies $F(\mathbf{u}_{\sigma 1}, \ldots, \mathbf{u}_{\sigma_n}) = (-1)^{\operatorname{sgn}\sigma} F(\mathbf{u}_1, \ldots, \mathbf{u}_n)$ for any permutation σ.
  - Now we return to the formula for $\det A$, which as we said can be written

$$\det A = \sum_{\sigma\in\Sigma_n}^{n} A_{\sigma 1,1}\cdots A_{\sigma n,n}\det(\mathbf{e}_{\sigma 1}, \ldots, \mathbf{e}_{\sigma n})$$

    which by the last observation is equal to

$$\sum_{\sigma\in\Sigma_n}^{n} (-1)^{\operatorname{sgn}\sigma} A_{\sigma 1,1}\cdots A_{\sigma n,n}\det(\mathbf{e}_1, \ldots, \mathbf{e}_n)$$

    which, using that $\det \mathrm{I} = 1$ is equal to

$$\sum_{\sigma\in\Sigma_n}^{n} (-1)^{\operatorname{sgn}\sigma} A_{\sigma 1,1}\cdots A_{\sigma n,n}.$$

- We conclude that any alternating multilinear function det with $\det \mathrm{I} = 1$ must be given by this explicit formula, which shows that there can be at most one such function.
- Moreover, if we *define* det by the above formula, then one can check that it does indeed satisfied the required properties, and we thus conclude not only the uniqueness, but the existence of the determinant.

- In the case of (2×2) or (3×3)-matrices, the above of course recovers the familiar formulas for the determinant.
- If we look at the last stage of the proof, we see we get a slightly more general statement: any function $F \colon K^{n \times n} \to K$ which is alternating and multilinear must be given by $F(A) = F(\mathrm{I}) \cdot \det(A)$.

22 Apr 21, More 3.6: Eigenvalues and eigenvectors

22.1 More 3.6B: Bases of eigenvectors

Properties of the determinant

- A useful feature of the explicit formula for the determinant is that it immediately reveals that $\det A = \det A^\top$ (as is seen by re-indexing the sum by replacing σ with σ^{-1}); hence the determinant is also the unique function which is multilinear and alternating with respect to the *rows* of a matrix.
- We recall some further important properties of determinant.
- The simplest case in which to compute a determinant is that of a *diagonal* matrix D, with diagonal entries $\lambda_1, \dots, \lambda_n$.
  - In this case, the determinant is just the product $\prod_{i=1}^n \lambda_i$.
  - This follows immediately from multilinearity and $\det \mathrm{I} = 1$.
  - Similarly, if A is (upper- or lower-) *triangular*, meaning that $A_{ij} = 0$ for $i < j$ (or for $i > j$)), then $\det A$ is again just the product of the diagonal entries.
  - This can be seen, for example, from the permutation formula, by noting that all but one of the terms in the sum will be zero; alternatively, it follows easily by induction using the Laplace expansion, which we turn to next.
- Next, a useful way to compute determinants it the **Laplace expansion** or **cofactor expansion** along a given row or column of a matrix $A \in K^{n\times n}$.
  - Here, for each i, j, we define the **minor** $M_{ij} \in K^{(n-1)\times(n-1)}$ of A to be the matrix obtained by removing the i-th row and j-th column from A.
  - The Laplace expansion along the j-th column is then the formula
$$\det A = \sum_{i=1}^{n} (-1)^{i+j} A_{ij} \det M_{ij}.$$
  - To prove it, write $\mathbf{v}_1, \dots, \mathbf{v}_n$ for the columns of A. Using linearity in the j-th column, we have $\det A = \sum_{i=1}^n A_{ij} \det(\mathbf{v}_1, \dots, \mathbf{v}_{j-1}, \mathbf{e}_i, \mathbf{v}_{j+1}, \dots, \mathbf{v}_n)$.
  - Thus, it remains to prove that $\det(\mathbf{v}_1, \dots, \mathbf{v}_{j-1}, \mathbf{e}_i, \mathbf{v}_{j+1}, \dots, \mathbf{v}_n) = (-1)^{i+j} \det M_{ij}$; we note immediately that since det is alternating, this reduces to showing left-hand side equals
$$(-1)^{n-i} \det(\mathbf{v}_1, \dots, \mathbf{v}_{j-1}, \mathbf{v}_{j+1}, \dots, \mathbf{v}_n, \mathbf{e}_i).$$
  - Because the determinant is alternating and multilinear, it follows that it is unchanged upon adding a multiple of one column to another column.
  - Hence $\det(\mathbf{v}_1, \dots, \mathbf{v}_{j-1}, \mathbf{v}_{j+1}, \dots, \mathbf{v}_n, \mathbf{e}_i)$ remains unchanged if we replace each $\mathbf{v}_k$ with $\overline{\mathbf{v}}_k = \mathbf{v}_k - A_{ik}\mathbf{e}_i$, i.e., if we replace the i-th entry of each $\mathbf{v}_k$ by 0.
  - Let us write $\mathbf{u}_1, \dots, \mathbf{u}_{n-1} \in K^{n-1}$ for the columns of M_{ij}, and let us write $\tilde{\mathbf{u}}_k \in K^n$ for the vector obtained from $\mathbf{u}_k$ by inserting a 0 as the i-th entry.
  - We note that the sequence $\overline{\mathbf{v}}_1, \dots, \overline{\mathbf{v}}_{j-1}, \overline{\mathbf{v}}_{j+1}, \dots \overline{\mathbf{v}}_n$ is precisely equal to $\tilde{\mathbf{u}}_1, \dots, \tilde{\mathbf{u}}_{n-1}$.

– We have thus reduced to showing that $(-1)^{i+j} \det M_{ij}$ is equal to $(-1)^{n-j} \det(\tilde{\mathbf{u}}_1, \dots, \tilde{\mathbf{u}}_{n-1}, \mathbf{e}_i)$, i.e., that $\det M_{ij} = (-1)^{n-i} \det(\tilde{\mathbf{u}}_1, \dots, \tilde{\mathbf{u}}_{n-1}, \mathbf{e}_i)$.

– Using that the determinant is also alternating with respect to *rows*, we have that the right-hand side is equal to $\det(\hat{\mathbf{u}}_1, \dots, \hat{\mathbf{u}}_{n-1}, \mathbf{e}_n)$, where $\hat{\mathbf{u}}_k \in K^n$ is the vector obtained from $\mathbf{u}_k \in K^{n-1}$ by adding a 0 at the *end*.

– We now claim in general for any vectors $\mathbf{w}_1, \dots, \mathbf{w}_{n-1} \in K^{n-1}$ that $\det(\mathbf{w}_1, \dots, \mathbf{w}_{n-1}) = \det(\hat{\mathbf{w}}_1, \dots, \hat{\mathbf{w}}_{n-1}, \mathbf{e}_n)$.

– Indeed, this follows immediately from the characterization of the determinant, since both sides are alternating, multilinear functions of $\mathbf{w}_1, \dots, \mathbf{w}_{n-1}$ which take the value 1 on $\mathbf{e}_1, \dots, \mathbf{e}_{n-1}$.

- Next, the most important property of the determinant is its **multiplicativity**, meaning $\det(AB) = \det A \cdot \det B$ for any $A, B \in K^{n \times n}$.

  – To prove this, fix A and consider the function $F(B) = \det(AB)$.

  – If B has columns $\mathbf{v}_1, \dots, \mathbf{v}_n$, then AB has columns $A\mathbf{v}_1, \dots, A\mathbf{v}_n$.

  – It is thus immediate that F is alternating, and also, since A is linear, that F is multilinear.

  – It follows that $F(B) = F(\mathrm{I}) \cdot \det(B) = \det(A \cdot \mathrm{I}) \det(B) = \det(A) \det(B)$ as desired.

- Another crucial property follows from this, which is that **if A is invertible, then** $\det A \neq 0$.

  – This follows since $\det A \det A^{-1} = \det(AA^{-1}) = \det \mathrm{I} = 1$.

  – Conversely, if A is *not* invertible, then $\det A = 0$.

    * To see this, we use that if A is not invertible, then we can write some column of A as a linear combination of the other columns, say $\det \mathbf{v}_n = \sum_{i=1}^{n-1} c_i \mathbf{v}_i$.
    * It follows using the alternating property that

$$\det A = \sum_{i=1}^{n-1} c_i \det(\mathbf{v}_1, \dots, \mathbf{v}_{n-1}, \mathbf{v}_i) = 0.$$

  – There is a second approach, which proves the contrapositive – if $\det A \neq 0$, then A is invertible – by giving an explicit formula the A^{-1}, based on the Laplace expansion.

    * Namely, with the notation M_{ij} as in the definition of the Laplace expansion, we have $(A^{-1})_{ij} = (\det A)^{-1}(-1)^{i+j} \det M_{ji}$.
    * Indeed, if we define A^{-1} in this way, then using the Laplace expansion formula, it follows that $(A^{-1}A)_{jj} = (\det A)^{-1} \sum_{i=1}^{n} (-1)^{i+j} A_{ij} \det M_{ij} = 1$, and when $j \neq k$, we have $(A^{-1}A)_{jk} = (\det A)^{-1} \sum_{i=1}^{n} (-1)^{i+j} A_{ij} \det M_{ik}$, which is 0, since $\sum_{i=1}^{n} (-1)^{i+j} A_{ij} \det M_{ik}$ is equal to the determinant of the matrix obtained by replacing the k-th column of A with the j-th column (and any matrix with two equal columns has zero determinant).

- Finally, one more important characterization of the determinant (which however we will not use, nor even formulate precisely) is that the $\det A$ is the (signed) *n-dimensional volume* of the parallelepiped $\{\sum_{i=1}^{n} a_i \mathbf{v}_i \mid 0 \leq a_1, \dots, a_n \leq 1\}$ spanned by the columns $\mathbf{v}_1, \dots, \mathbf{v}_n$ of A; more generally, for any subset $U \subset \mathbb{R}^n$ with a well-defined n-dimensional volume $\mathrm{vol}(U)$, we have $\mathrm{vol}(A(U)) = |\det A| \cdot \mathrm{vol}(U)$, where $A(U)$ is the image of U under the linear transformation A.

Characteristic polynomials

- Returning to eigenvalues, we said above that λ is an eigenvalue of $A \in K^{n\times n}$ if and only if $\lambda \mathrm{I} - A$ is not invertible, which we can now say is equivalent to $\det(\lambda \mathrm{I} - A) = 0$.
- The explicit formula for the determinant says that $\det(\lambda \mathrm{I} - A)$ is a sum of products of n entries of $\lambda \mathrm{I} - A$, where each such product contains one entry from each row, and in particular can contain at most n diagonal entries.
- Since all occurrences of λ are on the diagonal, it follows that this is a polynomial of degree n in λ.
- We call the resulting polynomial $p_A(x) = \det(x\mathrm{I} - A) \in K[x]$ the characteristic polynomial of A.
- It follows that λ is an eigenvalue of A if and only if $p_A(\lambda) = 0$.
- In particular, we see that an $(n \times n)$-matrix A has **at most** n **eigenvalues**, and in fact has *exactly* n (complex) eigenvalues, if counted with multiplicity.
- *Finding* the eigenvalues of A then just amounts to finding the roots of the polynomial $p_A(x)$.
- Once this is done, given an eigenvalue λ, finding the *eigenvectors* with eigenvalue λ then amounts to computing the nullspace of $\lambda I - A$ (or equivalently $A - \lambda I$), which we know how to do – this is just a matter of solving the homogeneous system of linear equations $(A - \lambda \mathrm{I})\mathbf{x} = \mathbf{0}$.
- As an example, a diagonal or triangular matrix with diagonal entries $\lambda_1, \ldots, \lambda_n$ has characteristic polynomial $\prod_{i=1}^{n}(x - \lambda_i)$ and hence the eigenvalues are exactly $\lambda_1, \ldots, \lambda_n$.
- Note in general that $\det(A) = (-1)^n \det(0 \cdot \mathrm{I} - A) = p_A(0) = (-1)^n \prod_{i=1}^{n}(0 - \lambda_i) = \prod_{i=1}^{n} \lambda_i$, i.e., the determinant is always the product of the eigenvalues (taken with multiplicities).
- Another important property of the characteristic polynomial is that it is *invariant under conjugation*: i.e., for any invertible matrix U, we have $p_{UAU^{-1}}(x) = p_A(x)$.
  - Indeed, we have:

$$\begin{aligned} p_{UAU^{-1}}(x) &= \det(x\mathrm{I} - UAU^{-1}) \\ &= \det(UxIU^{-1} - UAU^{-1}) \\ &= \det(U(x\mathrm{I} - A)U^{-1}) \\ &= \det(U)\det(x\mathrm{I} - A)\det(U)^{-1} \\ &= \det(x\mathrm{I} - A) \\ &= p_A(x). \end{aligned}$$

  - Hence, when we conjugate a matrix, the resulting matrix has the same eigenvalues with the same multiplicities.
  - This is consistent with the fact that we already know, that when we diagonalize a matrix A, the resulting diagonal entries are precisely the eigenvalues of A.

Example 3.6.3

- Returning to the above matrix $A = \begin{bmatrix} 1 & 1 \\ 4 & 1 \end{bmatrix}$, we have

$$xI - A = \begin{bmatrix} x-1 & -1 \\ -4 & x-1 \end{bmatrix}$$

and hence

$$p_A(x) = (x-1)^2 - 4 = x^2 - 2x - 3 = (x-3)(x+1)$$

with roots $\lambda_1 = 3$ and $\lambda_2 = -1$. These are thus the eigenvalue of A.

- To find the eigenvectors for λ_1 and λ_2, we solve

$$\begin{bmatrix} 3-1 & -1 \\ -4 & 3-1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \text{ and } \begin{bmatrix} -1-1 & -1 \\ -4 & -1-1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

.

- These reduce to $2x - y = 0$ and $-2x - y = 0$, with solutions $(1, 2)$ and $(1, -2)$, respectively.

Theorem 3.6.6: independence of eigenvectors

- We have seen that we can diagonalize any operator having a basis of eigenvectors. We now want to investigate under what conditions this occurs.
- **Theorem 3.6.6**: if $\mathbf{u}_1, \ldots, \mathbf{u}_k$ are eigenvectors of a linear operator L, and the corresponding eigenvalues $\lambda_1, \ldots, \lambda_k$ are all distinct, then the vectors $\mathbf{u}_i$ are linearly independent.
  - The proof is by induction on k, the base case $k = 1$ being trivial.
  - For the induction step, assume $k > 1$ and that $\mathbf{u}_1, \ldots, \mathbf{u}_{k-1}$ are linearly independent.
  - Now suppose that $c_1\mathbf{u}_1 + \cdots + c_k\mathbf{u}_k = \mathbf{0}$; we want to show $c_1 = \cdots = c_k = 0$.
  - Applying L to both sides, we obtain $c_1\lambda_1\mathbf{u}_1 + \cdots + c_k\lambda_k\mathbf{u}_{k-1} = \mathbf{0}$.
  - Subtracting λ_k times the first equation from the second, we obtain $c_1(\lambda_1 - \lambda_k)\mathbf{u}_1 + \cdots + c_{k-1}(\lambda_{k-1} - \lambda_k)\mathbf{u}_k = \mathbf{0}$.
  - By the induction hypothesis and since $\lambda_i \neq \lambda_k$ for all $i < k$, we conclude that $c_i = 0$ for all $i < k$, and hence that $c_k = 0$ as well, as required.
- We conclude that **if an $(n \times n)$-matrix has n distinct eigenvalues** – or in other words **if its characteristic polynomial has n distinct roots** – **then it is diagonalizable**.

Example 3.6.7

- Applying this theorem to the differentiation operator $D \in \mathcal{C}^\infty(\mathbb{R})$, we see that the functions e^{rx} for differing values of r are all linearly independent, which is something we had previously deduced from the uniqueness theorem for solutions to linear differential equations with constant coefficients.

Example 3.6.9

- The diagonal matrix $\operatorname{diag}(2,2,3)$ with diagonal entries $(2,2,3)$ is obviously diagonalizable, though it does not have 3 distinct eigenvalues.
- Hence, the condition in the above theorem is sufficient, but not necessary.
- By contrast, the matrix
$$A = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$
really isn't diagonalizable.
  - The reason is that, if it were to have a basis of eigenvectors, these would all have to have eigenvalue 2 or 3. But by computing the kernels of $A - 2\mathrm{I}$ and $A - 3\mathrm{I}$, we find that the only vectors with eigenvalue 2 and 3 are multiples of $\mathbf{e}_1$ and $\mathbf{e}_3$, respectively.
- As we mentioned previously, there are also examples like $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$, which is not diagonalizable as a *real* matrix (i.e., there is no *real* U with $U^{-1}AU$ diagonal) since both eigenvalues are complex, however it has two distinct *complex* eigenvalues $\pm i$, hence is diagonalizable as a complex matrix.

22.2 Triangulability

- As we have seen, it is **not** the case that every linear operator $L \in \mathcal{L}(V)$ on a finite-dimensional vector space V is diagonalizable.
- However, over the complex numbers, there are two nice results that do hold for every linear operator.
- The first says that every linear operator L on a finite-dimensional complex vector space V is triangulable.
- This means that V admits a basis $\mathcal{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_n)$ such that the matrix $A = [L]_{\mathcal{B}}$ is *upper-triangular* (i.e., $A_{ij} = 0$ for $j < i$).
  - Another way of saying this is that each $L\mathbf{b}_i$ is a linear combination only of itself and the earlier $\mathbf{b}_i$'s: $L\mathbf{b}_j = \sum_{i=1}^{j} A_{ij}\mathbf{b}_i$.
- The proof is by induction on the dimension n of V, the base case $n = 1$ being trivial.
  - For the induction step, suppose $n > 1$.
  - Choose an eigenvector $\mathbf{b}_1 \in V$ of L, say with $L\mathbf{b}_1 = \lambda\mathbf{b}_1$ (we know that L must have at least one eigenvector, since its characteristic polynomial has at least one root!).
  - Now complete $\mathbf{b}_1$ to a basis $\mathbf{b}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ of V, and set $W = \operatorname{Span}(\mathbf{v}_2, \ldots, \mathbf{v}_n) \subset V$.
  - We have a linear map $\Pi\colon V \to W$ defined by $\Pi\,\mathbf{b}_1 = \mathbf{0}$ and $\Pi\,\mathbf{v}_i = \mathbf{v}_i$ for $i > 1$.
  - By the induction hypothesis, the linear map $\Pi \circ L\colon W \to W$ is triangulable, so there is a basis $\mathbf{b}_2, \ldots, \mathbf{b}_n$ of W such that
$$\Pi(L\mathbf{b}_j) = \sum_{i=2}^{j} B_{ij}\mathbf{b}_i$$
for each $j \geq 2$, for some coefficients $B_{ij} \in \mathbb{C}$.

- It is easy to see that $\mathbf{b}_1, \ldots, \mathbf{b}_n$ is a basis of V; we claim that L is triangular with respect to this basis.
- Indeed, we have $L\mathbf{b}_1 = \lambda \mathbf{b}_1$, and by the definition of Π and the above equation, it follows for each j that $L\mathbf{b}_j = B_{1j}\mathbf{b}_1 + \sum_{i=2}^{j} B_{ij}\mathbf{b}_i$ for some $B_{1j} \in \mathbb{C}$, as desired.

- A restatement of this theorem is that for every square matrix $A \in \mathbb{C}^{n \times n}$, there is an invertible matrix U such that $U^{-1}AU$ is upper-triangular.

23 Apr 23, 13.1: Eigenvalues and eigenvectors

23.1 Jordan normal form

- We saw that, though not every complex square matrix is diagonalizable, they are all *triangulable*.
- We now explain a second result, which is a bit more complicated, but also much stronger: it says that each operator L on a finite-dimensional complex vector space V can be triangulated in a particularly simple form, called a *Jordan normal form*, and moreover that this representation is essentially *unique*.
  - First, we define a Jordan block to be a matrix of the form

$$\begin{bmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{bmatrix}$$

    for some $\lambda \in \mathbb{C}$ (all entries not shown are zero). In other words, it is equal to $\lambda \mathrm{I} + N$ where N is the matrix with

$$N_{ij} = \delta_{i+1,j} = \begin{cases} 1 & j = i+1 \\ 0 & \text{otherwise.} \end{cases}$$

  - A matrix is then in Jordan normal form if it is a block-diagonal matrix

$$\begin{bmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_k \end{bmatrix}$$

    where each block A_i is a Jordan block.

    * Formally, we can define a block diagonal matrix as follows: supposing that each A_i is a square matrix of size n_i and setting $n = \sum_{i=1}^{k} n_i$, we note that each $a \in \{1, \dots, n\}$ is of the form $a = \sum_{j=1}^{u_a - 1} n_j + v_a$ for some uniquely determined $u_a \in \{1, \dots, k\}$ and $v_a \in \{1, \dots, n_{u_a}\}$. Then the block-diagonal matrix A with blocks $A_1, \dots, A_k$ is defined by setting

$$A_{ab} = \begin{cases} (A_{u_a})_{v_a, v_b} & u_a = u_b \\ 0 & \text{otherwise} \end{cases}$$

    * This definition is hard to work with in practice. It is better to work with the following more abstract definition (which one should show is equivalent to the previous one): we say that a vector space V is a direct sum of the subspaces $W_1, \dots, W_k$ (denoted $V = W_1 \oplus \cdots \oplus W_k$) if for each $\mathbf{v} \in V$, there are unique $\mathbf{w}_i \in W_i$ for $i = 1, \dots, k$ such that $\mathbf{v} = \mathbf{w}_1, \dots, \mathbf{w}_k$. It is easy to see that, given ordered bases $\mathcal{B}_i$ of W_i for $i = 1, \dots, k$, the sequence of vectors $\mathcal{B} = \mathcal{B}_1 \cdots \mathcal{B}_k$ in V obtained by concatenating them is an ordered basis of V. Now the fact is that for an operator $L \in \mathcal{L}(V)$ on a finite-dimensional space V and an ordered basis $\mathcal{B}$ of V, the matrix $A = [L]_{\mathcal{B}}$ is block-diagonal with blocks $A_1, \dots, A_k$ if and only if there is a direct sum decomposition $V = W_1 \oplus \cdots \oplus W_k$ such that each W_i is invariant under L (i.e., $L\mathbf{w} \in W_i$ for $\mathbf{w} \in W_i$), and bases $\mathcal{B}_i$ of each W_i such that $\mathcal{B} = \mathcal{B}_1 \cdots \mathcal{B}_k$ and $A_i = [L|_{W_i}]_{\mathcal{B}_i}$. (In particular, a given matrix $A \in K^{n \times n}$ is block

diagonal if and only if the standard basis $\mathcal{E}$ of K^n can be decomposed into subsequences $\mathcal{E} = \mathcal{E}_1 \dots \mathcal{E}_k$ such that each $W_i = \operatorname{Span} \mathcal{E}_i$ is invariant under L; and in this case the i-th block is the matrix $A_i = [A|_{W_i}]_{\mathcal{E}_i}$.)

- The **theorem** (which is quite non-trivial, and I will not prove it here) is then that for every linear operator L on a finite-dimensional vector, there exists some basis $\mathcal{B}$ such that the matrix $[L]_{\mathcal{B}}$ is in Jordan normal form.
  - Again, an equivalent statement is that for every square matrix $A \in \mathbb{C}^{n \times n}$, there is an invertible matrix U such that $U^{-1}AU$ is in Jordan normal form.
  - As mentioned, there is a *uniqueness* statement as well: any two Jordan normal forms for L are the same, up to rearranging the Jordan blocks.
- Note: the letter N is used because this matrix is nilpotent meaning that some power of it is zero; in fact N^d is zero where $N \in \mathbb{C}^{n \times n}$.
  - More generally, we can compute N^k explicitly for any $k \geq 0$: it is given by $(N^k)_{ij} = \delta_{(i+k),j}$.
  - This is because $N\mathbf{e}_i = \mathbf{e}_{i-1}$ for $i = 1, \dots, d$ (where we set $\mathbf{e}_i = 0$ for $i \leq 0$) and hence $N^k \mathbf{e}_i = \mathbf{e}_{i-k}$.

23.2 13.1 Eigenvalues and eigenvectors

- We return to the linear system $\dot{\mathbf{x}} = A\mathbf{x}$ for some $A \in K^{n \times n}$.
- As we have seen, this has a solution $\mathbf{x} = e^{\lambda t}\mathbf{u}$ whenever $\mathbf{u}$ is an eigenvector of A with eigenvalue λ.

Example 13.1.1

- Consider the linear system with matrix $A = \begin{bmatrix} 1 & 1 \\ 4 & 1 \end{bmatrix}$.
- We have seen that A has eigenvectors $\mathbf{u}_1 = (1, 2)$ and $\mathbf{u}_2 = (1, -2)$ with eigenvalues 3 and -1, respectively.
- Hence, we obtain solutions
$$\mathbf{x}(t) = c_1 e^{3t} \begin{bmatrix} 1 \\ 2 \end{bmatrix} + c_2 e^{-t} \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$
for $c_1, c_2 \in K$.
- Moreover, for any initial condition $\mathbf{x}_0 \in K^2$, we can write $\mathbf{x}_0 = c_1\mathbf{u}_1 + c_2\mathbf{u}_2$ for some $c_1, c_2 \in K$, hence the solution $\mathbf{x} = c_1 e^{3t}\mathbf{u}_1 + c_2 e^{-t}\mathbf{u}_2$ satisfies $\mathbf{x}(0) = \mathbf{x}_0$.
- Since by the existence and uniqueness theorem, there is a *unique* solution with this initial condition, it follows that every solution is of the form $\mathbf{x}$ as above; i.e., this is the *general solution.*

13.1B: Eigenvector matrices

- More generally, the above argument shows that if A is diagonalizable with eigenbasis $\mathbf{u}_1, \dots, \mathbf{u}_n$ and eigenvalues $\lambda_1, \dots, \lambda_n$, then the general solution to $\dot{\mathbf{x}} = A\mathbf{x}$ is $\sum_{i=1}^n c_i e^{\lambda_i t}\mathbf{u}_i$.
- We can summarize this in a nice way by introducing the matrix U with columns $\mathbf{u}_1, \dots, \mathbf{u}_n$. Recall that we have seen that $A = UDU^{-1}$, where $D = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$ is the diagonal matrix with entries $(\lambda_1, \dots, \lambda_n)$.
- Now introduce the diagonal matrix $\Lambda_t = \mathrm{diag}(e^{\lambda_1 t}, \dots, e^{\lambda_n t})$.
- The general solution can then be written as $\mathbf{x} = U\Lambda_t\mathbf{c}$ with $\mathbf{c} = (c_1, \dots, c_n) \in K^n$ an arbitrary vector.
- We can see directly that this is a solution: we have that $\dot{\Lambda}_t = D\Lambda_t$ (where the derivative of a matrix-valued function is, as usual, defined entry-wise differentiation).
  - Next, since matrix multiplication satisfies the product rule $\frac{\mathrm{d}}{\mathrm{d}t}(XY) = \dot{X}Y + X\dot{Y}$ (this follows from the formula for matrix multiplication, and the product and sum rules for differentiation), and since the matrices U and $\mathbf{c}$ are constant, it follows that $\dot{x} = UD\Lambda_t\mathbf{c}$.
  - But we have $UD = AU$, hence $\dot{x} = AU\Lambda_t\mathbf{c} = A\mathbf{x}$, as required.
- We can similarly see directly that this is the general solution: given any initial condition $\mathbf{x}_0$, to find a solution $\mathbf{x} = U\Lambda_t\mathbf{c}$ with $\mathbf{x}(0) = \mathbf{x}_0$, we see we must have $\mathbf{x}_0 = U\Lambda_0\mathbf{c} = U\mathbf{c}$ and hence $\mathbf{c} = U^{-1}\mathbf{x}_0$. Thus, we see that $\mathbf{x} = U\Lambda_t U^{-1}\mathbf{x}_0$ is the desired solution.
  - Soon, we will see that, using the matrix exponential, this form of the solution form generalizes to the case where A is not diagonalizable.

Geometric interpretation

- Continuing with the assumption that A is diagonalizable, $A = UDU^{-1}$, we can rewrite the equation $\dot{\mathbf{x}} = A\mathbf{x}$ as $U^{-1}\dot{\mathbf{x}} = DU^{-1}\mathbf{x}$.
- Let us set $\mathbf{y} = U^{-1}\mathbf{x}$. Then $\mathbf{y} = [\mathbf{x}]_{\mathcal{B}}$ is just the coordinate vector of $\mathbf{x}$ with respect to the basis $\mathcal{B}$ consisting of the columns of U.
- The equation in $\mathbf{y}$ is then $\dot{\mathbf{y}} = D\mathbf{y}$.
- This is just an uncoupled equation with general solution $\mathbf{y} = \Lambda_t\mathbf{c}$. (This is the same general solution $\mathbf{x} = U\mathbf{y} = U\Lambda_t\mathbf{c}$ we obtained above.)
- In other words, if we use the coordinate system given by the eigenbasis $\mathcal{B}$, in which A is just given by scaling along the coordinate axes, then the differential equation becomes a simple, uncoupled equation.

24 Apr 28, 13.2: Matrix exponentials

24.1 13.2A: Definition

- For a square matrix A, we define its matrix exponential

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

  where by definition $A^0 = \mathrm{I}$.

  - As usual, this is meant as a limit of partial sums $\lim_{N\to\infty} \sum_{k=0}^{N} \frac{A^k}{k!}$, and where we are taking an entry-wise limit: i.e., for a sequence of matrices B_N, we say $B_N \xrightarrow{N\to\infty} B$ if $(B_N)_{ij} \xrightarrow{N\to\infty} B_{ij}$ for each (i,j).
  - Of course, a limit of a given sequence of matrices may or may not exist; and in particular, it is not *a priori* clear that the matrix exponential is always defined.

- As a simple example where the matrix exponential clearly *is* defined, consider a diagonal matrix $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$.

  - Then $D^k = \mathrm{diag}(\lambda_1^k, \ldots, \lambda_n^k)$, hence $\sum_{k=0}^{N} \frac{D^k}{k!} = \mathrm{diag}\left(\sum_{k=0}^{N} \frac{\lambda_1^k}{k!}, \ldots, \sum_{k=0}^{N} \frac{\lambda_n^k}{k!}\right)$ and hence in the limit we have that $e^D = \mathrm{diag}(e^{\lambda_1}, \ldots, e^{\lambda_n})$.
  - In particular, wit the notation Λ_t from the previous section, we see that $\Lambda_t = e^{tD}$.

Theorem 13.2.1

- **Theorem**: The matrix exponential e^A exists for all matrices $A \in \mathbb{C}^{n\times n}$.

  - Moreover, for a fixed A, the function $t \mapsto e^{tA}$ satisfies the following properties:
  - (a) $e^{(s+t)A} = e^{sA}e^{tA}$.
  - (b) $e^{tA}e^{-tA} = \mathrm{I}$.
  - (c) $\frac{\mathrm{d}}{\mathrm{d}t} e^{tA} = Ae^{tA} = e^{tA}A$.

- As a preliminary remark, note that for any two polynomials $p(x)$ and $q(x)$, the matrices $p(A)$ and $q(A)$ commute, i.e., $p(A) \cdot q(A) = q(A) \cdot p(A)$.

  - (We recall here that by $p(A)$ is defined as $\sum_{i=0}^{n} a_i A^i$, where $p(x) = \sum_{i=0}^{n} a_i x^i$, and $q(A)$ similarly.)
  - This is clear since A^m and A^n obviously commute for any m, n, and if a matrix X commutes with both Y and Z, then it also commutes with $aY + bZ$ for any scalars a and b. (More generally, if A and B commute, the so do $p(A)$ and $q(A)$.)
  - It follows that if $p(A) = \sum_{n=0}^{\infty} a_n A^n$ and $q(A) = \sum_{n=0}^{\infty} b_n A^n$ are convergent power series in A, then $p(A)$ commutes with $q(A)$ (and with $q(B)$ for any B commuting with A.)
  - Indeed, if we let $p_N(A)$ and $q_N(A)$ be the N-th partial sums, of these power series, then $p_N(A)q_N(A) = q_N(A)p_N(A)$.

- Now we may to the limit $N \to \infty$ to obtain $p(A)q(A) = q(A)p(A)$, since matrix multiplication commutes with limits: if $X_N \xrightarrow{N\to\infty} X$ and $Y_N \xrightarrow{N\to\infty} Y$ then $X_N \cdot Y_N \xrightarrow{N\to\infty} X \cdot Y$. This is because for a matrix the entries of a matrix product $X \cdot Y$ are all continuous functions in the entries of X and Y.

- Let us now see that e^A always exists.
  - That is, we must see that the sequence $\sum_{k=0}^{N} \frac{A^k}{k!}$ converges (entry-wise).
  - Choose some $b > 0$ such that $|A_{ij}| < b$ for all (i, j).
  - Since the entries of A^2 are of the form $\sum_{k=1}^{n} a_{ik}a_{kj}$, it follows from the triangle inequality that they are bounded in absolute value by nb^2.
  - Arguing similarly by induction, we find that the entries of A^k are bounded in absolute value by $n^{k-1}b^k$.
  - We may now apply the comparison test to each entry of e^A against the convergent series $1 + b + \sum_{k\geq 2}^{\infty} \frac{n^{k-1}b^k}{k!}$ to conclude that each entry of e^A is defined by an absolutely convergent series.

- We next turn to (c).
  - For the same reason as with ordinary power series, we can differentiate matrix-valued power series e^A term-by-term. (The reason is each entry of e^A is a so-called *uniform* limit of the functions defined by the partial sums.)
  - Hence, we have

$$\frac{\mathrm{d}}{\mathrm{d}t} e^{tA} = \frac{\mathrm{d}}{\mathrm{d}t} \sum_{k=0}^{\infty} \frac{t^k A^k}{k!} = \sum_{k=0}^{\infty} \frac{kt^{k-1}A^k}{k!} = \sum_{k=1}^{\infty} \frac{t^{k-1}A^k}{(k-1)!} = A \sum_{k=0}^{\infty} \frac{t^k A^k}{k!} = Ae^{tA},$$

    where in the last step, we use the same fact used above that we may interchange matrix multiplication with limits.

- We now prove (b).
  - Using the product rule for matrix multiplication, and the fact that A commutes with e^{sA} for any s, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(e^{-tA}e^{tA}\right) = -tAe^{-A}e^{tA} + Ae^{-tA}e^{tA} = 0.$$

  - It follows that each entry of $e^{-tA}e^{tA}$ is constant in t.
  - Since $e^{-0A}e^{0A} = \mathrm{I} \cdot \mathrm{I} = \mathrm{I}$, it follows that $e^{-tA}e^{tA} = \mathrm{I}$ for all t.

- Next, we prove that $f(t) = e^{tA}$ is the *unique* matrix-valued function satisfying $f'(t) = Af(t)$ and $f(0) = \mathrm{I}$.
  - This is analogous to the corresponding fact which we know about the ordinary exponential function, and the proof is the same:

- For any such function f, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\big(e^{-tA}f(t)\big) = -Ae^{-tA}f(t) + Ae^{-tA}f(t) = 0.$$

- Hence the entries of $e^{-tA}f(t)$ are all constant in t, hence since $e^{-0A}f(0) = \mathrm{I}\cdot\mathrm{I}$, it follows that $e^{-tA}f(t) = \mathrm{I}$ for all t.
- Hence $f(t) = (e^{-tA})^{-1} = e^{tA}$ for all t.

- Finally, we prove (a).
  - Fix $s \in \mathbb{R}$ and consider the function $g(t) = e^{-sA}e^{(t+s)A}$.
  - Using the product rule and the chain rule, we have $g'(t) = Ae^{-sA}e^{(t+s)A} = Ag(t)$.
  - We also have $g(0) = e^{-sA}e^{(0+s)A} = \mathrm{I}$.
  - Hence, by the above uniqueness statement, $g(t) = e^{tA}$, hence $e^{(t+s)A} = e^{sA}e^{tA}$, as desired.

24.2 13.2B: Solving systems

- Now consider the equation $\dot{\mathbf{x}} = A\mathbf{x}$ for any matrix A.
- Setting $\mathbf{x}(t) = e^{tA}\mathbf{u}$ for any $\mathbf{u} \in K^n$, we have using the product rule that $\dot{\mathbf{x}} = Ae^{tA}\mathbf{u} = A\mathbf{x}$, hence this is a solution.
- Moreover, this is the *general* solution, since for any initial condition $\mathbf{x}_0$, we have $e^{0A}\mathbf{x}_0 = \mathbf{x}_0$, hence $e^{tA}\mathbf{x}_0$ is the unique solution satisfying this initial condition.
- We can also see directly that it is the general solution, without appealing to the uniqueness theorem: if $\mathbf{x}$ is any solution, so that $\dot{x} = A\mathbf{x}$, we then have by the product rule $\frac{\mathrm{d}}{\mathrm{d}t}(e^{-At}\mathbf{x}) = -Ae^{-At}\mathbf{x} + Ae^{-At}\mathbf{x} = \mathbf{0}$, hence $e^{-At}\mathbf{x}$ must be equal to some constant $\mathbf{u}$, and hence $\mathbf{x} = e^{At}\mathbf{u}$.
- Note also that if $A\mathbf{u} = \lambda\mathbf{u}$, then (as you'll show in the homework) $A^n\mathbf{u} = \lambda^n\mathbf{u}$ for all n, and hence $e^{tA}\mathbf{u} = e^{\lambda t}\mathbf{u}$.
  - Hence, in this case the solution $e^{tA}\mathbf{u}$ becomes the solution $e^{\lambda t}\mathbf{u}$ we found previously (as it had to, by the uniqueness theorem).

24.3 13.2C: Relationship to eigenvectors

- If A is a diagonal matrix $A = D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, we have $e^{tA} = \Lambda_t$ in the notation from above, so in this case, we recover the general solution $\mathbf{x} = \Lambda_t\mathbf{u}$ to the uncoupled system $\dot{\mathbf{x}} = A\mathbf{x}$.
- Now suppose more generally that A is *diagonalizable* so that $A = UDU^{-1}$.
- We now make the crucial observation that exponentiation commutes with conjugation: $e^{YXY^{-1}} = Ye^XY^{-1}$. Indeed, we have

$$\begin{aligned} e^{YXY^{-1}} &= \lim_{N\to\infty}\sum_{k=0}^{N}\frac{(YXY^{-1})^k}{k!} = \lim_{N\to\infty}\sum_{k=0}^{N}\frac{YX^kY^{-1}}{k!} \\ &= \lim_{N\to\infty} Y\Big(\sum_{k=0}^{N}\frac{X^k}{k!}\Big)Y^{-1} = Y\Big(\lim_{N\to\infty}\sum_{k=0}^{N}\frac{X^k}{k!}\Big)Y^{-1} = Ye^XY^{-1}, \end{aligned}$$

where in the penultimate step we are once again using the continuity of matrix multiplication, and where in the second step, we are using that $(YAY^{-1})(YBY^{-1}) = Y(A\cdot B)Y^{-1}$ for any $A, B \in \mathbb{C}^{n\times n}$, and similarly, by induction, for the product of the conjugates of any number of matrices.

- Hence, the general solution to $\dot{\mathbf{x}} = A\mathbf{x}$ is

$$e^{tA}\mathbf{u} = e^{U(tD)U^{-1}}\mathbf{u} = Ue^{tD}U^{-1}\mathbf{u} = U\Lambda_t U^{-1}\mathbf{u},$$

recovering the general solution we found before.

- Let us now look at a *non*-diagonalizable example.

Example 13.2.1

- Let $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$.
- Note that $A^n = \begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix}$ and hence $(tA)^n = \begin{bmatrix} t^n & kt^n \\ 0 & t^n \end{bmatrix}$.
- Noting that $\sum_{k=0}^{\infty} k\frac{t^k}{k!} = t\frac{\mathrm{d}}{\mathrm{d}t}e^t = te^t$, it follows that

$$e^{tA} = \begin{bmatrix} e^t & te^t \\ 0 & e^t \end{bmatrix}.$$

- We thus have the general solution

$$\mathbf{x} = e^t A \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} c_1 e^t + c_2 te^t \\ c_2 e^t \end{bmatrix}.$$

- We thus see the reappearance of the term te^t that came up in linear equations with repeated roots.
  - Note here, that $p_A(x)$ does indeed have the repeated root 1 with multiplicity 2.
  - But beware! The matrix $A = \mathrm{I}$ *also* has 1 as a double root, but *its* general solution is $\mathbf{x} = (c_1 e^t, c_2 e^t)$.
  - The difference is that in the second case, the multiplicity 2 eigenvalue corresponds to two linearly independent eigenvectors, whereas in the first case, there is only one eigenvector

25 Apr 30, 13.2D: Computing matrix exponentials

25.1 13.2D: Computing e^{tA} in practice.

- We know how to compute e^{tA} when A is diagonalizable (by diagonalizing A), and we saw how to compute e^{tA} in the above simple example. We now seek how to compute it in general.
- This can be done by using the Jordan normal form $A = UJU^{-1}$ of A.
  - Recall that $(tA)^k = U(tJ)^kU^{-1}$ and $e^{tk} = Ue^{tJ}U^{-1}$.
  - Moreover, products and sums of block-diagonal matrices (with blocks of the same size and in the same order) are computed block-wise
    - I.e., if X and Y are block-diagonal with blocks $X_1, \dots, X_r$ and $Y_1, \dots, Y_r$, respectively, with X_i of the same size as Y_i, then $X + Y$ and $X \cdot Y$ are block-diagonal with blocks $X_1 + Y_1, \dots, X_r + Y_r$ and $X_1 \cdot Y_1, \dots, X_r \cdot Y_r$, respectively.
    - (This is because if $V = W_1 \oplus \cdots \oplus W_k$ is a direct-sum decomposition of a vector space V and each W_i is invariant with respect to both $L_1, L_2 \in \mathcal{L}(V)$, then each W_i is also invariant with respect to $L_1 + L_2$ and $L_1 \circ L_2$.)
  - Hence if J has Jordan blocks $J_1, \dots, J_s$, then e^{tJ} has blocks $e^{tJ_1}, \dots, e^{tJ_s}$.
  - Hence, if we can compute the exponential for a single Jordan block, we can compute it for the entire Jordan normal form.
- Now, recall that each Jordan block J_i has the form $J_i = \lambda I + N$, where N is the matrix given by $N_{ij} = \delta_{(i+1),j}$. Let d be the size of the Jordan block J_i.
  - Using that $N^j = 0$ for $j \geq d$ (more generally, it is easy to see that N^j is the matrix with zeros everywhere, and ones on the diagonal which is j entries above the main diagonal), it follows that
$$(tJ_i)^k = t^k \sum_{j=0}^{k} \binom{k}{j} \lambda^{k-j} N^j = t^k \sum_{j=0}^{d-1} \binom{k}{j} \lambda^{k-j} N^j$$
  - Hence
$$\begin{aligned} e^{tJ_i} &= \sum_{k=0}^{\infty} \frac{t^k}{k!} \sum_{j=0}^{d-1} \binom{k}{j} \lambda^{k-j} N^j = \sum_{j=0}^{d-1} N^j \sum_{k=j}^{\infty} \frac{t^k}{k!} \binom{k}{j} \lambda^{k-j} \\ &= \sum_{j=0}^{d-1} \frac{N^j}{j!} \sum_{k=j}^{\infty} \frac{t^k}{(k-j)!} \lambda^{k-j} = \sum_{j=0}^{d-1} \frac{N^j}{j!} \sum_{k=0}^{\infty} \frac{t^{k+j}}{k!} \lambda^k = \sum_{j=0}^{d-1} \frac{t^j e^{\lambda t}}{j!} N^j \end{aligned}$$
  - Hence, e^{tJ_i} looks like
$$e^{\lambda t} \cdot \begin{bmatrix} 1 & t & \frac{t^2}{2} & \frac{t^3}{3!} & \cdots & \frac{t^{d-1}}{(d-1)!} \\ 0 & 1 & t & \frac{t^2}{2} & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & \frac{t^2}{2} \\ & & & & 1 & t \\ 0 & & \cdots & & 0 & 1 \end{bmatrix}.$$

- We again see the reappearance of the terms $t^k e^{\lambda t}$ we saw for linear differential equations with multiple roots; but again, note here that it is not just the *multiplicity of the roots* that matters, but the *size of the Jordan blocks*.

- We conclude that e^{tJ} is a block-diagonal matrix with blocks e^{tJ_i} as above.
- Hence, finally, $e^{tA} = Ue^{tJ}U^{-1}$, and hence we have an explicit way to compute e^{tA} once we have its Jordan normal form.
  - Of course, this only helps if we *can* find the Jordan normal form of A – i.e., find the basis U such that $U^{-1}AU$ is in Jordan normal form.
  - There is a general algorithm to do this as well, once the eigenvalues of A are known (which we will not present).
  - We next present an alternative method to compute e^{tA} which is often easier.
  - But first, let us work out an example directly using the Jordan normal form.

Example 13.2.5 using Jordan normal form

- Consider the matrix $A = \begin{bmatrix} 0 & 1 \\ -4 & -4 \end{bmatrix}$.
- This has characteristic polynomial $\det \begin{bmatrix} \lambda & -1 \\ 4 & \lambda + 4 \end{bmatrix} = \lambda^2 + 4\lambda + 4 = (\lambda + 2)^2$ with a double root $\lambda = -2$.
- Hence, the only two possibilities for the Jordan normal form are $\begin{bmatrix} -2 & 0 \\ 0 & -2 \end{bmatrix}$ and $\begin{bmatrix} -2 & 1 \\ 0 & -2 \end{bmatrix}$.
- Next, let us find the eigenvectors: we have $\ker(A - (-2)\mathrm{I}) = \ker \begin{bmatrix} 2 & 1 \\ -4 & -2 \end{bmatrix} = \mathrm{Span}\left\{ \begin{bmatrix} 1 \\ -2 \end{bmatrix} \right\}$.
- We see that A has only one independent eigenvector, so it cannot be diagonalizable. Thus, its Jordan normal form is $J = \begin{bmatrix} -2 & 1 \\ 0 & -2 \end{bmatrix}$.
- Next, we seek the matrix U such that $A = UJU^{-1}$, or in other words, the basis $\mathbf{u}_1, \mathbf{u}_2$ such that J is the matrix for A with respect to this basis.
- Recalling what it means for J to represent A with respect to the basis $\mathbf{u}_1, \mathbf{u}_2$, and inspecting J, we see that we must have $A\mathbf{u}_1 = -2\mathbf{u}_1$, hence $\mathbf{u}_1$ is the (unique up to scaling) eigenvector $\mathbf{u}_1 = (1, -2)$; and we see that $A\mathbf{u}_2 = \mathbf{u}_1 - 2\mathbf{u}_2$, or in other words $(A - 2\mathrm{I})\mathbf{u}_2 = \mathbf{u}_1$.
- Thus to find $\mathbf{u}_2 = (x, y)$, we solve $\begin{bmatrix} 2 & 1 \\ -4 & -2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$, and we find $(x, y) = (1, -1)$.
- In conclusion, we have $U = \begin{bmatrix} 1 & 1 \\ -2 & -1 \end{bmatrix}$, and hence $U^{-1} = \begin{bmatrix} -1 & -1 \\ 2 & 1 \end{bmatrix}$.
  - (As a consistency test, we can verify that $A = UJU^{-1}$ as required.)
- Hence:

$$e^{tA} = Ue^{tJ}U^{-1} = e^{-2t}U \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix} U^{-1} = e^{-2t} \begin{bmatrix} 1 + 2t & t \\ -4t & 1 - 2t \end{bmatrix}.$$

26 May 5, More 13.2D: Computing matrix exponentials with Cayley-Hamilton

26.1 13.2.5 Cayley-Hamilton theorem

- We next turn to the promised alternative method to compute the matrix exponential.
  - We will prove the following **Theorem**: e^{tA} can be written as a linear combination $e^{tA} = \sum_{j=0}^{n-1} b_j(t)A^j$ for some smooth functions $b_j(t)$.
  - The proof will make use of the *Cayley-Hamilton theorem*.
- The **Cayley-Hamilton theorem** says that $p_A(A) = 0$ for any $A \in \mathbb{C}^{n\times n}$, where p_A is the characteristic polynomial of A.
  - This follows easily from the existence of the Jordan normal form $A = UJU^{-1}$.
  - Indeed, if A has eigenvalues $\lambda_1, \ldots, \lambda_s$ with multiplicities $d_1, \ldots, d_s$, then $p_A(x) = \prod_{i=1}^{s}(x - \lambda_i)^{d_i}$ and $p_A(A) = \prod_{i=1}^{n}(A - \lambda_i \mathrm{I})^{d_i} = U\prod_{i=1}^{n}(J - \lambda_i \mathrm{I})^{d_i}U^{-1}$.
  - But now each Jordan block J_k is of the form $\lambda_i \mathrm{I}+N$ for some i and has size at most $\ell \leq d_i$ (since the Jordan block J_k contributes ℓ copies of the eigenvalue λ_k to the characteristic polynomial).
  - Hence $(J_k - \lambda_i \mathrm{I})^{d_i} = N^{d_i} = 0$.
  - Hence $\prod_{i=1}^{s}(J_k - \lambda_i I)^{d_i} = 0$ for each k, and since polynomials of block-diagonal matrices are evaluated block-wise, it follows that $\prod_{i=1}^{s}(J - \lambda_i I)^{d_i} = 0$.
- Note that this proof actually gives a stronger result, namely that $m_A(A) = 0$ for a certain polynomial $m_A(x)$ called the minimal polynomial, which is *smaller than*, i.e., *divides* $p_A(x)$.
  - $m_A(x)$ is defined by $m_A(x) = \prod_{i=1}^{s}(x - \lambda_i)^{m_i}$, where m_i is the size of the largest Jordan block of A with eigenvalue λ_i.
  - It is called *minimal* because, as is easily proven, it is the unique monic polynomial of least degree with $m_A(A) = 0$, or also because it divides every polynomial p with $p(A) = 0$.
  - It will follow that in the theorem we are working towards – i.e., that e^{tA} is a combination of the powers of A – we similarly only need to take the first d powers, where d is the degree of $m_A(x)$.
- Let us now prove that e^{tA} can be written as a linear combination $e^{tA} = \sum_{j=0}^{d-1} b_j(t)A^j$ for some smooth functions $b_j(t)$, with d the degree of $m_A(x)$.
- To begin with, it follows immediately from the Cayley-Hamilton theorem that A^d is a linear combination $\sum_{j=0}^{d-1} c_j A^j$.
- By induction, it follows that for all $N \geq 0$, A^N is a linear combination of $A^0, \ldots, A^{d-1}$, i.e., we have $A^N = c_0^{(N)}A^0 + \cdots + c_{d-1}^{(N)}A^{d-1}$ for some coefficients $c_0^{(N)}, \ldots, c_{d-1}^{(N)}$.
- Specifically, we find that for $N \geq d$, the coefficients $c_j^{(N)}$ satisfy the recurrence relations $c_0^{(N+1)} = c_0 c_{d-1}^{(N)}$ and $c_j^{(N+1)} = c_j c_{d-1}^{(N)} + c_{j-1}^{(N)}$ for $j > 0$.

- We also have the base case $c_j^{(N)} = \delta_{jN} c_j$ for $N < d$.
- It follows by induction that if $|c_j| < b$ for all j, and $b > 1$, then $|c_j^{(N)}| < (2b)^N$ for all N.

- We conclude that
$$e^{tA} = \sum_{k=0}^{\infty} t^k \frac{A^k}{k!} = \sum_{k=0}^{\infty} t^k \sum_{j=0}^{d-1} \frac{c_j^{(k)}}{k!} A^j = \sum_{j=0}^{d-1} A^j \sum_{k=0}^{\infty} \frac{c_j^{(k)}}{k!} t^k,$$
as desired, where the series $b_j(t) = \sum_{k=0}^{\infty} \frac{c_j^{(k)}}{k!} t^k$ converges absolutely by the above estimates, and where the possibility of exchanging the two sums also follows from absolute convergence.

- Thus, once we compute the powers $A^0, A^1, A^2, \ldots, A^{d-1}$, computing e^{tA} boils down to finding the coefficient functions $b_j(t)$, which we turn to next.

Example 13.2.8: side remark: Cayley Hamilton and inverses

- They Cayley-Hamilton theorem gives an additional method for computing the inverse of a matrix.
- Namely, if $p_A(x) = \sum_{i=0}^{n} a_i x^i$, then $0 = p_A(A) = \sum_{i=0}^{n} a_i A^i$.
- If A is invertible, then $a_0 = \det A \neq 0$, and we can then solve for $A^0 = \mathrm{I}$ to obtain
$$\mathrm{I} = A^0 = A(-a_0 \sum_{i=1}^{n-1} a_i A^{i-1}).$$
- This shows that $A^{-1} = -a_0 \sum_{i=1}^{n-1} a_i A^{i-1}$.
- Example: let's compute the inverse of $A = \begin{bmatrix} 2 & 3 & 1 \\ 0 & 2 & 2 \\ 0 & 0 & 2 \end{bmatrix}$ this way.
- We have $p_A(x) = (x-2)^3 = x^3 - 6x^2 + 12x - 8$.
  - Hence $0 = p_A(A) = A^3 - 6A^2 + 12A - 8$.
  - Solving for A^{-1}, we obtain:
$$A^{-1} = \frac{1}{8}(A^2 - 6A + 12\mathrm{I})$$
- Next, we compute
$$A^2 = \begin{bmatrix} 4 & 12 & 10 \\ 0 & 4 & 8 \\ 0 & 0 & 4 \end{bmatrix}$$
- Hence
$$A^{-1} = \frac{1}{8}(A^2 - 6A + 12\mathrm{I}) = \frac{1}{8} \begin{bmatrix} 4 & -6 & 4 \\ 0 & 4 & -4 \\ 0 & 0 & 4 \end{bmatrix}.$$

Theorem 13.2.4

- **Theorem**: We can write e^{At} as a linear combination $e^{At} = \sum_{j=0}^{n-1} b_j(t)A^j$ in which the coefficient functions $b_j(t)$ satisfy

$$e^{t\lambda_k} = \sum_{j=0}^{n-1} b_j(t)\lambda_k^j \tag{$\mathrm{E}_{\lambda_k,0}$}$$

  for all eigenvalues $\lambda_1, \ldots, \lambda_r$. Moreover, if some λ_k appears with multiplicity m, the $b_j(t)$ can be chosen so as to also satisfy

$$\left.\frac{\mathrm{d}^i}{\mathrm{d}\lambda^i}\right|_{\lambda=\lambda_k} e^{t\lambda} = \left.\frac{\mathrm{d}^i}{\mathrm{d}\lambda^i}\right|_{\lambda=\lambda_k} \sum_{j=0}^{n-1} b_j(t)\lambda^j \tag{$\mathrm{E}_{\lambda_k,i}$}$$

  for each $i = 1, \ldots, m-1$.

- As a first remark, note that the above equations ($\mathrm{E}_{\lambda_k,i}$) are a system of n linear equations in the functions $b_j(t)$.

  - In the case when all the eigenvalues are distinct, the matrix of coefficients appearing in this system is exactly the Vandermonde matrix, and is therefore invertible. Hence, we see that there is a unique choice of functions $b_0, \ldots, b_{n-1}$ satisfying these equations.

  - Similarly, in the case where some eigenvalues appear with multiplicities, the resulting matrix is the generalized Vandermonde matrix which came up in the proof of Theorem 11.2.4, which as we saw there is also invertible. Hence, in this case too, there is a unique choice of $b_0, \ldots, b_{n-1}$ satisfying the equations.

  - (An important consequence of this is that the coefficients $b_0(t), \ldots, b_{n-1}(t)$ in the formula $e^{tA} = \sum_{j=0}^{n-1} b_j(t)A^j$ for the exponential *only depend on the eigenvalues* of A.)

- Now, we first prove a part of the theorem. Namely, we prove that any functions $b_0, \ldots, b_{n-1}$ with $e^{tA} = \sum_{i=0}^{n-1} b_i A^i$ must satisfy the equations ($\mathrm{E}_{\lambda_k,i}$) for $i = 0, 1, \ldots, m'-1$, where $m' \leq m$ is the size of the largest Jordan block of A with eigenvalue λ_k. (Note that the number of such equations is exactly the degree d of the minimal polynomial of A.)

  - By definition, we have $e^{tA} = \sum_{j=0}^{n-1} b_j(t)A^j$.

  - Taking the Jordan normal form $A = UJU^{-1}$, it follows by conjugating both sides that $e^{tJ} = \sum_{j=0}^{n-1} b_j(t)J^j$.

  - Now, fix any λ_k, and let $B = \lambda_k \mathrm{I} + N$ be the largest Jordan block of J with eigenvalue λ_k, so that the size of B is m' as above (recall that N is the square matrix with zeros everywhere and ones just above the diagonal: $N_{ab} = \delta_{(a+1),b}$).

  - Since multiplication of block-diagonal matrices is compute block-wise, we conclude that

$$e^{tB} = \sum_{j=0}^{n-1} b_j(t)B^j.$$

- Now, we know from an earlier computation that the left-hand side of the above equation $e^{tB} = e^{\lambda t}\sum_{i=0}^{m'-1} \frac{t^i}{i!}N^i$.

- Let us similarly compute the right-hand side.

– Since $B = \lambda_k \mathrm{I} + N$, we have $B^j = \sum_{i=0}^{j} \binom{j}{i} \lambda_k^{j-i} N^i$.

– Hence:

$$\sum_{j=0}^{n-1} b_j B^j = \sum_{j=0}^{n-1} b_j \sum_{i=0}^{j} \binom{j}{i} \lambda_k^{j-i} N^i = \sum_{i=0}^{n-1} N^i \sum_{j=i}^{n-1} \binom{j}{i} b_j \lambda_k^{j-i} = \sum_{i=0}^{m-1} N^i \sum_{j=i}^{n-1} \binom{j}{i} b_j \lambda_k^{j-i}$$

- Comparing the two sides and noting that the N^i are linearly independent for $i = 0, \ldots, m' - 1$ (this follows by explicitly computing the powers of $N \in \mathbb{C}^{m' \times m'}$), we find that

$$e^{\lambda_k t} \frac{t^i}{i!} = \sum_{j=i}^{n-1} \binom{j}{i} b_j \lambda_k^{j-i}$$

for each $i = 0, \ldots, m - 1$.

– Setting $i = 0$, this gives the first equation

$$e^{\lambda t} = \sum_{j=0}^{n-1} b_j \lambda_k^j$$

– For the case $i > 0$, we multiply both sides by $i!$ to obtain

$$e^{\lambda_k t} t^i = \sum_{j=i}^{n-1} \frac{j!}{(j-i)!} b_j \lambda_k^{j-i}.$$

– But then we see that the left-hand side is

$$\left. \frac{\mathrm{d}^i}{\mathrm{d}\lambda^i} \right|_{\lambda = \lambda_k} e^{t\lambda}$$

and the right hand side is

$$\left. \frac{\mathrm{d}^i}{\mathrm{d}\lambda^i} \right|_{\lambda = \lambda_k} \sum_{j=0}^{n-1} b_j(t) \lambda^j$$

as desired.

- We have shown that any $b_0, \ldots, b_{n-1}$ with $e^{tA} = \sum_{j=0}^{n-1} b_j A^j$ must satisfy $(\mathrm{E}_{\lambda_k, i})$ with $0 \leq i < m'$. Let us call the set of these equations $\mathcal{E}'$. It remains to show that the b_j can be so chosen as to also satisfy $(\mathrm{E}_{\lambda_k, i})$ with $m' \leq i < m$. Let us write $\mathcal{E}$ for the entire system of equations. Hence $\mathcal{E}$ is a system of n equations, and $\mathcal{E}'$ is a system of d equations, where d is the degree of the minimal polynomial of A.

– Now, by the remark above about the invertibility of the coefficient matrix of the system $\mathcal{E}$, we know that there is a unique solutions $\tilde{b}_0, \ldots, \tilde{b}_{n-1}$ to the entire system $\mathcal{E}$.

– Since, by the Cayley-Hamilton argument we gave previously, each A^l with $l \geq d$, as well as e^{tA}, is a linear combination of $A^0, \ldots, A^{d-1}$, it follows that there exist functions $b_0, \ldots, b_{d-1}$ satisfying

$$e^{tA} - \sum_{j=d}^{n-1} \tilde{b}_j(t) A^j = \sum_{j=0}^{d-1} b_j(t) A^j. \qquad (\star)$$

– Let us now consider $\mathcal{E}'$ to be a system of equations in $b_0, \ldots, b_{d-1}$, by fixing the values of the remaining variables $b_d, \ldots, b_{n-1}$ to be $\tilde{b}_d, \ldots, \tilde{b}_{n-1}$.

– By the first part of the proof, we know that any $b_0, \ldots, b_{d-1}$ satisfying $(\star)$ must also satisfy $\mathcal{E}'$. We also know that $\tilde{b}_0, \ldots, \tilde{b}_{d-1}$ satisfy $\mathcal{E}'$.

– But the coefficients matrix of $\mathcal{E}'$ is a $d \times d$ (generalized) Vandermonde matrix, and is hence invertible, and so we conclude that $b_j = \tilde{b}_j$ for $j = 0, \ldots, d-1$.

– Hence, finally, we have that $\tilde{b}_0, \ldots, \tilde{b}_{n-1}$ satisfy both $\mathcal{E}$ and $e^{tA} = \sum_{j=0}^{n-1} \tilde{b}_j A^j$, as desired.

Example 13.2.5 again

- Let us now repeat Example 13.2.5 but now using this theorem. We have $A = \begin{bmatrix} 0 & 1 \\ -4 & -4 \end{bmatrix}$.
- We know that $e^{tA} = b_0(t)\mathrm{I} + b_1(t)A$ for some functions b_0 and b_1.
- Moreover, since we have a single eigenvalue $\lambda_1 = -2$ with multiplicity 2, the theorem says that we have the equations

$$e^{\lambda_1 t} = b_0(t)\lambda_1^0 + b_1(t)\lambda_1^1 \qquad \frac{\mathrm{d}}{\mathrm{d}\lambda}\bigg|_{\lambda=\lambda_1} e^{\lambda t} = \frac{\mathrm{d}}{\mathrm{d}\lambda}\bigg|_{\lambda=\lambda_1} b_0(t)\lambda^0 + b_1(t)\lambda^1,$$

  i.e.,

$$e^{-2t} = b_0(t) - 2b_1(t) \qquad te^{-2t} = b_1(t).$$

- We thus have $b_1 = te^{-2t}$ and hence $b_0 = e^{-2t} + 2te^{-2t}$.
- We conclude that

$$e^{tA} = b_0\mathrm{I} + b_1 A = e^{-2t}\begin{bmatrix} 1+2t & t \\ -4t & 1-2t \end{bmatrix},$$

  which is the same answer we got before.

27 May 7, 13.3: Nonhomogeneous systems

27.1 13.3A: Solution formula

- We can handle non-homogeneous linear systems of ODEs just in the same way we handled a single non-homogeneous linear equation, using exponential multipliers.
- First, let us recall the general linear-algebra fact that given any particular solution $\mathbf{x_p}$ to the equation

$$\dot{\mathbf{x}}(t) = A\mathbf{x}(t) + \mathbf{b}(t),$$

the *general solution* will be

$$\mathbf{x} = \mathbf{x}_\mathrm{p} + \mathbf{x}_\mathrm{h},$$

where $\mathbf{x}_\mathrm{h}$ is a general *homogeneous* solution, i.e., a general solution to

$$\dot{\mathbf{x}}(t) = A\mathbf{x}(t).$$

- Thus, since we already know how to solve the homogeneous equation, finding the general solution to the inhomogeneous equation amounts to finding any one particular solution.

Theorem 3.2

- For any $t_0 \in \mathbb{R}$, a particular solution $\mathbf{x}_\mathrm{p}$ to

$$\dot{\mathbf{x}}(t) = A\mathbf{x}(t) + \mathbf{b}(t)$$

is given by

$$\mathbf{x}_\mathrm{p}(t) = e^{tA} \int_{t_0}^{t} e^{-\tau A}\mathbf{b}(\tau)\,\mathrm{d}\tau$$

- Remarks:
  - From the just-mentioned general principle about inhomogeneous equations, it follows that the *general* solution is

$$\mathbf{x} = \mathbf{x}_\mathrm{p} + \mathbf{x}_\mathrm{h} = e^{tA} \int_{t_0}^{t} e^{-\tau A}\mathbf{b}(\tau)\,\mathrm{d}\tau + e^{tA}\mathbf{c}$$

    with $\mathbf{c} \in K^n$.
  - Note that if we change t_0, this will simply have the effect of modifying the constant $\mathbf{c}$.
  - More generally, if we replace $\int_{t_0}^{t} e^{-\tau A}\mathbf{b}(\tau)\,\mathrm{d}\tau$ with any *antiderivative* of $e^{-tA}\mathbf{b}(t)$, we will obtain a general solution.
    - Unlike in the single-variable case, the *general* antiderivative of a vector-valued function $\mathbf{y}(t)$ is not $\int_{t_0}^{t} \mathbf{y}(\tau)\,\mathrm{d}\tau$ but rather $\left(\int_{t_1}^{t} y_1(\tau)\,\mathrm{d}\tau, \ldots, \int_{t_n}^{t} y_n(\tau)\,\mathrm{d}\tau\right)$, i.e., we can choose a different constant of integration on each component.
    - Hence, concretely, instead of $\int_{t_0}^{t} e^{-\tau A}\mathbf{b}(\tau)\,\mathrm{d}\tau$, we can just choose an antiderivative of each component of $e^{-tA}\mathbf{b}(t)$.
    - Note that in the book, they denote this by $\int e^{-tA}\mathbf{b}(t)\,\mathrm{d}t$, but beware that, as we just said, this does not correspond to a single definite integral, but rather a separate definite integral on each component.

* Finally, note that if we take a general anti-derivative $\int e^{-tA}\mathbf{b}(t)\,\mathrm{d}t$ in this sense, the result $e^{tA}\int e^{-tA}\mathbf{b}(t)\,\mathrm{d}t$ is actually the *general* solution, since the varying choices of the constant $\mathbf{c}$ in the homogeneous term simply correspond to different choices of anti-derivatives.

- Proof:
  - As mentioned, the proof is just an application of exponential multipliers.
  - We multiply both sides of the equation $\mathbf{x} = A\mathbf{x} + \mathbf{b}$ by the invertible matrix e^{tA} to obtain the equivalent equation
$$e^{-tA}\mathbf{x} - e^{-tA}A\mathbf{x} = e^{tA}\mathbf{b},$$
which we can also write as
$$\frac{\mathrm{d}}{\mathrm{d}t}\big(e^{-tA}\mathbf{x}\big) = e^{-tA}\mathbf{b}.$$
  - Applying the fundamental theorem of calculus coordinate-wise, we find a particular solution to this equation by integrating from any t_0:
$$e^{-tA}\mathbf{x}_\mathrm{p}(t) = \int_{t_0}^{t} e^{-\tau A}\mathbf{b}(\tau)\,\mathrm{d}\tau.$$
  - We then solve for $\mathbf{x}_\mathrm{p}$:
$$\mathbf{x}_\mathrm{p} = e^{tA}\int_{t_0}^{t} e^{-\tau A}\mathbf{b}(\tau)\,\mathrm{d}\tau.$$
  - As mentioned above, in the penultimate step, we could have instead chosen an arbitrary antiderivative "$\int e^{-tA}\mathbf{b}(t)\,\mathrm{d}t$", and would then arrive instead at the general solution $\mathbf{x}(t) = e^{tA}\int e^{-tA}\mathbf{b}(t)\,\mathrm{d}t$.

Example 13.3.1

- Let's solve
$$\dot{\mathbf{x}} = A\mathbf{x} + \mathbf{b} = \begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix}\mathbf{x} + \begin{bmatrix}e^t\\ e^{-t}\end{bmatrix}.$$
- We have already considered the corresponding homogeneous equation: since A is in Jordan normal form, we have
$$e^{At} = \begin{bmatrix}e^t & te^t\\ 0 & e^t\end{bmatrix}.$$
and hence the general homogeneous solution is
$$\mathbf{x}_\mathrm{h} = e^{At}\mathbf{c} = \begin{bmatrix}c_1e^t + c_2te^t\\ c_2e^t\end{bmatrix}.$$
- We now compute a particular inhomogeneous solution.

– We have

$$\begin{aligned}
\mathbf{x}_{\mathrm{p}} &= e^{tA} \int e^{-tA} \mathbf{b}(t)\, \mathrm{d}t \\
&= e^{tA} \int \begin{bmatrix} e^{-t} & -te^{-t} \\ 0 & e^{-t} \end{bmatrix} \begin{bmatrix} e^{t} \\ e^{-t} \end{bmatrix} \mathrm{d}t \\
&= e^{tA} \int \begin{bmatrix} 1 - te^{-2t} \\ e^{-2t} \end{bmatrix} \mathrm{d}t \\
&= e^{tA} \begin{bmatrix} t + \frac{1}{2} te^{-2t} + \frac{1}{4} e^{-2t} \\ -\frac{1}{2} e^{-2t} \end{bmatrix} \\
&= \begin{bmatrix} te^{t} + \frac{1}{4} e^{-t} \\ -\frac{1}{2} e^{-t} \end{bmatrix}
\end{aligned}$$

- Hence, the general solution is

$$\mathbf{x} = \mathbf{x}_{\mathrm{p}} + \mathbf{x}_{\mathrm{h}} = \begin{bmatrix} te^{t} + \frac{1}{4} e^{-t} + c_1 e^{t} + c_2 te^{t} \\ -\frac{1}{2} e^{-t} + c_2 e^{t} \end{bmatrix}.$$