1 January 8

1.1 Logistics

Within the three body problem is the entirety of this course.

"In mathematics you don't understand things. You just get used to them." - John von Neumann. (Bro von Neumann is so washed for this.) For context, von Neumann was challenged that he didn't actually understand the method of characteristics. He was talking about this with his Dad, who just retired as a math prof from Ohio State, and he didn't like it. Joel's alternate form: "In mathematics it takes time for ideas and examples to sink in."

Click here for course website link.

Content: Parts of Chapters 10-12, 14-16 of textbook by Ascher and Greif (online, free). Topics: Interpolation, approximation (Ch 10 - 12); Differentiation, Integration, ODE's [PDE's] (Ch 14-16). Back in the day, 303 was meant as a followup to 302, but not anymore, so may be some repeated material (norms and condition numbers, but not thee point of this course anyway).

Discussion: please post to piazza page. If this fails, please email to jf@cs.ubc.ca with subject CPSC 303.

Grading: $(10\%)\max(h, m, f) + (35\%)\max(m, f) + (55\%)f$ where h is homework, m is midterm, f is final. So technically, can ace final and ace course. Because back in the day, Joel knew someone who couldn't attend any classes, but got 100% in the final only to get C+ in the course. There is a phenomena where if someone gets 100 on the midterm, stops doing homework. But really, these assessments are good preperation and indicators of where you stand in the course.

Please sign up for piaaza and gradescope through canvas.ubc.ca (especially gradescope).

Homework: Set Thursday 11:59pm and due Thursday 11:59pm. There is both individual and group homework. Group homework: at most 4 people, and will cover most material; only submit one. Individual homework: These are the types of things he wants to make sure everyone can do, and will be like the things on exams; you must write up your own solution even if you work with others.

1.2 Intro to ODE's

1.2 and 4.2 (norms), 14.2 (differentiation), 16.1 and 16.2 (ODE's).

He typically begins courses with the most difficult stuff the course will get. This is tough, but not the worst. Now, this course only requires two terms of calculus, but with how pertinent ML is now, most people are taking multivariable calc anyway. We will get some idea of what to expect and some intuition, but will revisit later in the course. The reason the emphasis is on ODEs and not PDEs is because the general theory for ODEs really applies to all of them, even if you have to solve differently. Families of PDEs have their own properties that have to be studied separately.

We are heading towards Ordinary Differential Equations (Ch. 16).

1.2.1 Absolute vs. Relative Error

If $v \in \mathbb{R}$ is an approximation to $u \in \mathbb{R}$, then absolute error (in v) (as an approximation to u) is $|u - v|$, and the relative error is $\frac{|u-v|}{|u|}$. The same works in $\mathbb{R}^n$ (or any normed vector space):

$$\|\vec{u}\|_2 = \|(u_1, \ldots, u_n)\|_2 = \sqrt{u_1^2 + u_2^2 + \cdots u_n^2}$$

We also use $\|\vec{u}\|_1 = |u_1| + \cdots |u_n|$ and $\|\vec{u}\|_{\max} = \|\vec{u}\|_\infty = \max_{1 \le i \le n} |u_i|$. The absolute error in $\vec{v}$ as an approximation to $\vec{u}$ is $\|\vec{u} - \vec{v}\|_p$ and the relative error is $\frac{\|\vec{u}-\vec{v}\|}{\|\vec{u}\|_p}$ where $p = 1, 2, \infty$.

1.2.2 Taylor's Theorem (p. 5)

Theorem 1. *For $f\colon (a, b) \to \mathbb{R}$ where f is $k+1$ differentiable (so $f^{(k+1)}(x)$ exists in (a, b)), for some x_0, $x_0 + h$ that lie in (a, b),*

$$f(x_0 + h) = f(x_0) + hf'(x_0) + \frac{h^2}{2} f''(x_0) + \cdots + \frac{h^k}{k!} f^{(k)}(x_0) + \text{error}$$

where the error is $\frac{h^{k+1}}{(k+1)!} f^{(k+1)}(\xi)$ where ξ is between x_0 and $x_0 + h$.

Remark 1. h can be negative.

We'll learn how to approximate derivatives, and then ODE solution.

2 January 10

Housekeeping:

- HW1 will be assigned on Han 11, due on gradescope on Jan 18
- Access gradscope via Canvas
- Today: Separable ODE's (see, e.g. CLP 2, Appendix D of Ascher and Grief)

Last time: Ch 1: "Reviewing" termonology. Ch 4: $\|\vec{u}\|_2 = \sqrt{u_1^2 + \cdots + u_n^2}$ for $\vec{u} \in \mathbb{R}^n$. We saw classes of functions, Taylor Series, and ODE's. An ODE is where the derivatives are a function of the same variable?? So $y' = f(t, y)$. PDEs on the other hand have partial derivatives $h = h(t, x_1, \ldots, x_n)$, and maybe something like the heat equation $\frac{\partial h}{\partial t} = -\Delta_{x_1,\ldots,x_n} h$. There might be a course on the foundations of elliptic PDEs, parabolic PDEs, etc., but won't be the focus here.

And then he talks about a sketch of proof of Taylor's, except I literally did it this morning. Essentially, we use MVT to get better approximations (not rigourously). We have

$$f(x_0 + h) = f(x_0) + hf'(x_0) + \frac{h^2}{2} f''(x_0) + \frac{h^3}{6} f'''(\xi_1)$$

$$hf'(x_0) = f(x_0 + h) - f(x_0) - \frac{h^2}{2} f''(\xi_2)$$

$$f'(x_0) = \frac{f(x_0 + h) - f(x_0)}{h} - \frac{h}{2} f''(\xi_2) = \frac{f(x_0 + h) - f(x_0)}{h} + \text{Order}(h)$$

where $O(h)$ is some function (depends on f, x_0). We have $|\text{Order}(h)\| \leq \frac{h}{2} M_2$ where M_2 is bound on f'' in the interval $[x_0, x_0 + h]$. By definition, $f'(x_0) = \lim_{h \to 0} \frac{f(x_0+h) - f(x_0)}{h}$. All this is really taken from [A&G], Ch. 14, Sections 1,2.

We will continue our discussion of derivatives, but first, some useful notation. If $a < b \in \mathbb{R}$, $C[a, b] := \{f \colon [a, b] \to \mathbb{R}$ such that f is continuous$\}$. When $k \in \mathbb{N}$,

$$C^k(a, b) = \{f \colon (a, b) \to \mathbb{R} \text{ such that } f \text{ has } k \text{ continuous derivatives } \forall x \in (a, b)\}$$

, and similarly for $C^k[a, b]$. $C^\infty(a, b)$ is the set of f that has derivatieves of all order.

Definition 1 (Real Analytic Functions).

$$C^\omega(a, b) := \{f \colon (a, b) \to \mathbb{R} \text{ such that for all } x_0 \in (a, b),\ f(x_0 + h) = f(x_0) + hf'(x_0) + \frac{h^2}{2!} f''(x_0) + \cdots\}$$

A function $f \colon (a, b) \to \mathbb{R}$ is real analytic when $f \in C^\omega(a, b)$.

Real analytic is a stronger condition than C^∞, so $C^\omega \subset C^\infty \subset \cdots \subset C^2 \subset C^1 \subset C^0$. [Hmm... I would like a better definition of convergence here ff]

For example, e^x ff he scrolled.

2.1 Start ODE

Simple ODE [A&G]:

$$y' = f(t, y)$$

where we use the notation $y' = \frac{dy}{dt} = \dot{y}$. (Caution: math textbooks typically use $y' = \frac{dy}{dx} = f(x, y)$.)

To solve for $y = y(t)$, we are given an "intial condition", that is we have $y_o, t_o \in \mathbb{R}$ and impose $y(t_0) = y_0$.

We expect a unique solution. Say we are given $A \in \mathbb{R}$, and find a y that satisfies

$$y'(t) = Ay(t)$$

We claim that $y(t) = e^{At}C$ is a solution. We can verify: $y'(t) = (e^{At}C)' = (Ae^{At})C = Ay(t)$. So we have solved it by shamelessly guessing.

We can plug in our t_0 and y_0 which fixes $C = y_0 e^{-At_0}$. Then

$$y(t) = y_0 e^{A(t-t_0)}$$

So when A is big enough (greater than 0), we have exponential growth.

Is this solution unique? Can we simply guess and it provides the only solution? More on Friday.

3 January 12

Today:

- ODE: $y' = y$ "isoclines"
- ODE's of the form $y' = f(y)$
- Later $\vec{y}' = \vec{f}(t, \vec{y})$, system of m ODE's where $\vec{y} = \vec{y}(t) \colon \mathbb{R} \to \mathbb{R}^m$ and $\vec{f}(t, y) \colon \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^m$.
- Examples; $y' = y^2$, $y' = |y|^{1/2}$.

Question, what could possibly go wrong?

Solve $y' = f(y)$, $y(t_0) = y_0$. Then

(a). If f continuous near $y = y_0$ then a solution exists locally

(b). Moreover, if f is Lipschitz (or differentiable) near $y = y_0$, the local solution is unique

(c). If f is analytic near $y = y_0$, the unique local solution is analytic

(d). If $|f(y)| \leq Ky$ for some K constant, then there is a global solution

We will start looking at Matlab next week: we will specifically be looking for how it breaks. He doesn't cover as much content as other people do in 303, but more in depth. "I just became a grandfather last week, so I get to say back in my day we always had to make our own software." So we always knew how it broke. Now we have to figure out why other people's code breaks.

Recall last time, we were looking at $y' = Ay$, $y(t_0) = y_0$ where $y \colon \mathbb{R} \to \mathbb{R}$. Isocline picture: slopes at unit points in the $y - t$ plane. And then we guessed an answer last time: $y(t) = e^{A(t-t_0)}y_0$. Now, if we had a theorem that said this was unique, we'd be done.

We can solve this with integration: $y' = y \implies \frac{dy}{y} = dt \implies \int \frac{1}{y} dy = \int dt$, and so $\ln(y) + c_1 = t + c_2$ hence $y = e^{t+c}$. Now with t_0, y_0, we can determine our constant to get $y(t) = e^{t-t_0}y$. However, this method is not foolproof. Our solution could blow up to infinity. Consider $y' = y^2$ and $y(1) = 1$. And then $\frac{1}{y^2}dy = dt \implies \frac{-1}{y} = t + c$. Then, $y = \frac{-1}{t+c}$, and plugging in the initial conditions, we get $c = -2$. So $y = \frac{1}{2-t}$. Singularities can happen, as $y(t) \to \infty$ as $t \to 2$.

What about $y' = |y|^{1/2}$? What could possible go wrong? When $y > 0$ we have $y' = y^{1/2}$. Solving in the way we did before, we can get $y^{1/2} = \frac{1}{2}(t + C)$ so $y(t) = \frac{1}{4}(t + C)^2$. Note $y(-C) = 0$... but our slope should always be increasing, yet is 0 at a point? ff idk Now when $y < 0$, we can find (with the method as before) that $y = -\frac{1}{4}(t+C)^2$. So we have piecewise function of y depending on if $y > 0$ or $y < 0$.

Is this a unique solution? Actually, let $a < b$. Then

$$y(t) = \begin{cases} \frac{1}{4}(t-b)^2 & b \leq t \\ 0 & a \leq t \leq b \\ -\frac{1}{4}(t-a)^2 & t \leq a \end{cases}$$

is also a solution. The bad situation occurs when $y = 0$. We could stay "arbitrarily" long at $y = 0$, even if to the right and left are parabola! We will continue looking at $y' = |y|^{1/2}$ next time.

4 January 17

Today's outline:

- Euler's method
- MATLAB and Euler's method
- Plugging in $y' = Ay, y(t_0) = y_0$

4.1 Euler's Method

Say we are given $y' = f(t, y)$ (or $y' = f(y)$, or $\vec{y}' = \vec{f}(t, \vec{y})$, etc.) where f is a function, and we have) initial value $y(t_0) = y_0$, $y_0, t_0 \in \mathbb{R}$. We have the approximation

$$y'(t) \approx \frac{y(t+h) - y(t)}{h}$$

for small h, since $y'(t) = \lim_{h\to 0} \frac{y(t+h)-y(t)}{h}$. But

$$y'(t) \approx \frac{y(t+h) - y(t-h)}{2h}$$

is a much better approximation (usually). Rearranging gives $y(t+h) \approx y(t) + hy'(t)$; in fact, by Taylor's theorem, $y(t+h) = y(t) + hy'(t) + \frac{h^2}{2}y''(\xi)$ for some ξ between t and $t+h$. Substituting f, we have

$$y(t+h) = y(t) + hf(t, y) + \frac{h^2}{2}y''(\xi)$$

So for small h,

$$y(t+h) \approx y(t) + hf(t, y)$$

(we will ignore the final term for now, but will be useful later for calculating error). This last approximation is the essence of Euler's method.

So given t_0 = initial time and y_0 = initial value,

Step 1. Start with $y(t_0) = y_0$. [Pick a value of h, smaller the better (usually)]

Step 2. $y(t_0 + h) = y_0 + hf(t_0, y(t_0)) := y_1$

Step 3. $y(t_0 + 2h) = y_1 + hf(t_1, y_1) := y_2$ where $t_i = t_{i-1} + h = ih + t_0$.

Step n. Repeat

Actual ODE solvers (like in MATLAB) change h based off of f (especially makes h small when f is very large or f is changing quickly). When we saw the three body problem, we had error when two bodies got close because the gravitational force got so big it couldn't make h small enough to figure out what was going on.

If $y' = 2y$ and given y_0, t_0, recall that the exact solution was $y(t) = e^{2(t-t_0)}$. With the numerical approximation, we get $y_1 = y_0 + h(2y_0) = y_0(1 + 2h)$ and $y_2 = y - 1 + h(2y_1) = (1 + 2h)y_1$. So $y(t_i) = y(t_0 + ih) = (1 + 2h)^i y_0$. Now say we fix a t_{end}, perhaps N steps and so $t_{\text{end}} = t_0 + Nh$. Then $h = \frac{t_{\text{end}}-t_0}{N}$. So

$$y(t_{\text{end}}) = (1 + 2h)^N y_0 = \left(1 + \frac{2(t_{\text{end}} - t_0)}{N}\right)^N y_0 = \left(1 + \frac{\text{something}}{N}\right)^N y_0$$

Hmm... this looks a lot like a defition of e. As $N \to \infty$, we have $e^{\text{something}} y_0 = e^{2(t_{\text{end}}-t_0} y_0$. We will see this in MATLAB. On the homework, we will see MATLAB not working too well for $y' = |y|^{1/2}$.

And then he shows us his MATLAB for Euler's method. MATLAB is quirky. The first function you define in a file will always be run when you call the file, and it will assume the other functions in the file are not meant to be run globally.

5 January 24

Topics to finish:

- Look at MATLAB for a bit (Solutions to HW2 and HW1-ish)
- Vandermonde matrices and linear algebra without linear algebra
- Exponentiation and eigenpairs and norms
- Higher ordeer versions of Euler's method
- More celestial mechanics

Some quirky MATLAB: Something good to know is putting a `;` at the end of a line to supress the output (so we don't see giant vector). `1e-200` is defined as its own thing, but `1e-400` is considered `0`. `1/0` gives `Inf` and `-1/0` gives `-Inf`. But `-Inf + Inf` gives `NaN` (not a number). But all this follows the IEEE standard. MATLAB also calls the first function in a file as the name of a file when called elsewhere. We will look more at MATLAB later when we try to break it.

5.1 Vandermonde matrices

Given

$$A = \begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{bmatrix}$$

Does this have a unique solution?

The following are equivalent:

1. $A \begin{bmatrix} c_0 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} y_0 \\ \vdots \\ y_n \end{bmatrix}$ has a unique solution for all $y_0, \dots, y_n$.
2. A is invertible
3. $\det(A) \neq 0$
4. A is of rank $n+1$

We have actually seen Vandermonde matrices in HW1, just disguised. We were looking at 3 point schemes. We took a bunch of these formulas:

$$\begin{cases} f(x_0 + L) = \cdots \\ f(x_0) = \cdots \\ f(x_0 - h) = \cdots \\ f(x_0 + 2h) = \cdots \\ f(x_0 + \sqrt{2}h) = \cdots \\ f(x_0 + rh) = \cdots \end{cases}$$

(where r is a real number, often an integer, often $0, 1, 2, -1, -2, \dots$), and used the formula

$$\begin{aligned} c_0 \cdot f(x_0 + r_0 h) &= c_0 \left(f(x_0) + r_0 j f'(x_0) + r_0^2 \frac{h^2}{2} f''(x_0) + r_0^3 \frac{h^3}{3!} f'''(x_0) + O(h^4) \right) \\ c_1 \cdot f(x_0 + r_1 h) &= c_1(\dots) \\ c_2 \cdot f(x_0 + r_1 h) &= c_2(\dots) \\ c_3 \cdot f(x_0 + r_1 h) &= c_3(\dots) \end{aligned}$$

and we try to rearrange and change c_i and derive a scheme to get $0f(x_0)+1hf'(x_0)+0\frac{h^2}{2}f''(x_0)+0\frac{h^3}{3!}f'''(x_0)$. I.e., we are looking for c_0, c_1, c_2, c_3 such that

$$\begin{aligned}0 &= c_0 + c_1 + c_2 + c_3\\ 1 &= r_0c_0 + r_1c_1 + r_2c_2 + r_2c_3\\ 0 &= r_0^2c_0 + r_1^2c_1 + r_2^2c_2 + r_2^2c_3\\ 0 &= r_0^3c_0 + r_1^3c_1 + r_2^3c_2 + r_2^3c_3\end{aligned}$$

In homework 1, we saw this was solving a vandermonde matrix (anything that is of the form of the matrix below):

$$\begin{bmatrix}1 & 1 & 1 & 1\\ r_0 & r_1 & r_2 & r_3\\ r_0^2 & r_1^2 & r_2^2 & r_3^2\\ r_0^3 & r_1^3 & r_2^3 & r_3^3\end{bmatrix}\begin{bmatrix}c_0\\c_1\\c_2\\c_3\end{bmatrix} = \begin{bmatrix}0\\1\\0\\0\end{bmatrix}$$

(Observation: this looks like interpolation.)

Theorem 2. *The above system has a unique solution.*

Lemma 1. *Say that $p\colon \mathbb{R}\to\mathbb{R}$, $p = p(x)$ is a polynomial and p has $n+1$ distincr roots, $x_0, \dots, x_n$. Then*

1. *$p(x) = (x-x_0)(x-x_1)\dots(x-x_n)q(x)$ where $q(x)$ is a polynomial. (Bruh people in this class are actually so stupid, none of them recognize factoring).*
2. *$p'(x)$ has n distinct roots between x_0 and x_n (assuming $x_0 < x_1 < \cdots < x_n$).*

(a). $p''(x)$ has $n-1$ distinct roots between x_0 and x_n

(b). $p'''(x)$ has $n-2$ distinct roots between x_0 and x_n

To prove all of these, just done by Rolle's theorem

Theorem 3 (Rolle's Theorem)**.** *If $f \in C^0[a,b]$ (continuous on $[a,b]$), $f(a) = f(b) = 0$, and f is differentiable on (a,b), then $f'(\xi) = 0$ for some $a < \xi < b$.*

(Bro why is no one putting up their hand to haveing SEEN Rolle's theorem before, I swear there are too many bums in this course.) "How many people have seen a proof of this? Yeah, you would do this in Math 320." (Loewen could never) But this is quite intuitive anyway.

We do get a useful fact from the lemma:

Corollary 1. *If $p = p(x)$ is a polynomial of degree $\le n$, $p(x) = c_0 + c_1x + \cdots + c_nx^n$ and p has $n+1$ distinct roots, then $p = 0$*

Remark 2. $p(x) = x^2 + 1$, then p has no real roots. But $p'(x) = 2x$ has a real root, so the lemma doesn't go the other way (i.e. $q(x) = 0$ from bullet point 1.).

Claim: the system

$$\begin{bmatrix}1 & x_0 & x_0^2 & \cdots & x_0^n\\ 1 & x_1 & x_1^2 & \cdots & x_1^n\\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_n & x_n^2 & \cdots & x_n^n\end{bmatrix}\begin{bmatrix}c_1\\c_2\\\vdots\\c_n\end{bmatrix} = 0$$

has $c_1 = c_2 = \cdots = c_n = 0$ as its unique solution. Why? This just says $p(x) = c_0 + c_1x + c_2x^2 + \cdots + c_nx^n$ satisfies $p(x_0) = 0, p(x_1) = 0, \dots, p(x_n) = 0$, so p has $n+1$ distinct roots, hence by our corollary, $p = 0$. (Important in the definition of a Vandermonde matrix is that $x_0, x_1, \dots, x_n$ are distinct, otherwise not invertible).

Homework: $e^{\begin{bmatrix}0 & -1\\ 1 & 0\end{bmatrix}t} = \begin{bmatrix}\cos t & -\sin t\\ \sin t & \cos t\end{bmatrix}$ and $e^{\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}t} = \begin{bmatrix}\cosh t & \sinh t\\ \sinh t & \cosh t\end{bmatrix}$ (meant to get to this today).

6 January 26

Coming soon (later, we're putting a pin in it):

- More ODE's:
  - Central force: $m\vec{x}'' = -\frac{\vec{x}}{|\vec{x}|}g(|\vec{x}|)$,
  - Newton's law: $g(r) = \frac{1}{r^2}$
  - ff
- ff

Today:

- Higher Order ODE solvers. For now, Euler's method and Explicit trapezoidal
- More on e^{At}, norms, convergence, etc.s
- Where do non-diagnolizable matrices ceom from? One place: recurrences

If $y' = f(t, y), y(t_0) = y_0$, we can rewrite this as

$$y(t) - y_0 = \int_{s=t_0}^{s=t} f(s, y)ds$$

This is the integral form of "$y' = f(t, y)$". This is actually a perferred form for many applications, since we don't know if y is differentiable so we don't assume it. Rename: $f(s) = f(s, y)$ (or are we assuming??) Then $y' - f(t), y(t_0) = y_0$, then $y(t) - y(t_0) = \int_{s=t_0}^{s=t} f(s)ds$. We want to use a trapezoid to approximate the area under the curve, since it is a better approximation than a rectangle.

We have $y(t_0 + h) = y(t_0) + hy'(t_0) + O(h^2)$. Let $t_i = t_0 + ih$. We want y_1 to approximation $y(t_1)$. $y(t_1) = y(t_0 + h) \approx y(t_0) + hy'(t_0) + O(h^2)$. So $y_{i+1} = y_i + hf(t_0, y_0)$. So y_{i+1} approximates $y(t_{i+1})$ (uses y_i, approximates $y(t_i)$).

How do we improve? $y(t + h) = y(t) + h\frac{y'(t)+y'(t+h)}{2}$, which is a better approximation for y'. Trapezoidal: $\frac{y(t+h)-y(t)}{2}$ to $y'(t + \frac{h}{2})$. (I have no clue what he is doing.)

So first, we have $y(t_0), h$. Using Euler, we get $y(t_1) = y(t_0 + h) \approx y(t_0) + hy'(t_0) = y(t_0) + hf(t_0, y_0)$. $y_1 \leftarrow y(t_0) + hf(t_0, y_0)$.

Trapezoidal (Explicit) Scheme: Let $Y = y(t_0) + hf(t_0, y_0)$. And $y(t_0 + h) \approx y(t_0) + h\left(\frac{f(t_0,y_0)+f(t_1,Y)}{2}\right)$. The Y is like the otherside of the trapezoid.

6.1 Pre-interpolation

Before we get to interpolation, we need to get back at matrices, and what norms of matrices are.

6.1.1 Matrix exponential

Question: What do we mean by

$$e^A = I + A + \frac{A^2}{2} + \cdots$$

If $\vec{x} \in \mathbb{R}^n$, then $\|\vec{x}\|_2$ or $\|\vec{x}\|_{L^2}$ is $\sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$ (I feel like this should be ℓ^2).

Definition 2 (Matrix norm L^2). If $M \in \mathbb{R}^{n\times n}$, i.e. an $n \times n$ matrix, real entries, then we define

$$\|m\|_{L^2} = \max_{\vec{x}\neq\vec{0}} \frac{\|M\vec{x}\|_2}{\|\vec{x}\|_2}$$

So $\|M\|_{L^2}$ is the smallest real B such that $\|M\vec{x}\|_{L^2} \leq \|\vec{x}\|_{L^2} B$.

Example: $\vec{x} \in \mathbb{R}^n$, $n = 1$, $\vec{x} = x_1 \in \mathbb{R}^1$ so $A \in \mathbb{R}^1$. Say $A = \begin{bmatrix} 3 \end{bmatrix}$. $\max\left(\frac{\|A\vec{x}\|_{L^2}}{\|\vec{x}\|_{L^2}}\right)$. In this case, $\vec{x} = (x_1)$, so $\|\vec{x}\|_{L^2} = \sqrt{x_1^2} = |x_1|$. So $A \in \mathbb{R}^{|x|}$??? what is this notation, $A = \begin{bmatrix} a \end{bmatrix}$, so $\|A\|_{L^2} = |a|$. ff

We also have other norms: $\|\vec{x}\|_{\max} = \|\vec{x}\|_\infty = \max(|x_1|, |x_2|, \ldots, |x_n|)$. $\|A\|_{L^\infty} = \max_{\vec{x} \neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty}$.

Now look at 2×2 matrices. Say

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

$\text{size}(A) = \max(|a_{11}|, |a_{12}|, |a_{21}|, |a_{22}|)$. Claim: $\|A\|_\infty = \|A\|_2 = \text{size}(A)$, and so it doesn't matter what norm we pick.

What is $I + A + \frac{A^2}{2} + \frac{A^3}{3!} + \cdots$. What does convergence mean? The rth term is $I + A + \frac{A^2}{2} + \cdots + \frac{A^r}{r!}$. So we want to measure convergence:

$$\lim_{s,r \to \infty, s > r} \left(\frac{A^{r+1}}{(r+1)!} + \cdots + \frac{A^s}{s!} \right) \to 0$$

(I'm assuming this is Cauchy). With a given metric, say 2, this means

$$\lim_{s,r \to \infty, s > r} \left\| \frac{A^{r+1}}{(r+1)!} + \cdots + \frac{A^s}{s!} \right\|_2 \to 0$$

(but it is the same with the other metrics). (I'm curious, how many of our properties from $\mathbb{R}^n$ carry over to these matrix spaces, like we always have equivalence of these L^2, L^∞ norms?). We have e^{At} where $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. Then computing our terms,

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$
$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$
$$A^2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$
$$A^3 = A^2 A = IA = A$$

So for every powers, it is I, and if it is an odd power, it is A. Then we have

$$e^{At} = I + At + \frac{(At)^2}{2!} + \frac{(At)^3}{3!} + \cdots = \begin{bmatrix} 1 + \frac{t^2}{2} + \frac{t^4}{4!} + \cdots & t + \frac{t^3}{3!} + \frac{t^5}{5!} + \cdots \\ t + \frac{t^3}{3!} + \frac{t^5}{5!} + \cdots & 1 + \frac{t^2}{2} + \frac{t^4}{4!} + \cdots \end{bmatrix} = \begin{bmatrix} \cosh(x) & \sinh(x) \\ \sinh(x) & \cosh(x) \end{bmatrix}$$

We'll do more diagonalization things next time.

7 January 29

Today:

- e^{At} where $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$
- Other reasons to write $A = S \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} S^{-1}$ other than e^{At}, namely recurrences and $A^n, n = 0, 1, \ldots$.
- Examples of
  (a). Diagonlizable A
  (b). Defective A (in recurrences) ("not enough eigenvectors")... (not diagonlizable, because it does not have a complete basis of eigenvectors, some duplicates)

What is a 2×2 defective matrix? Means a (real) matrix that is not diagonliazble over $\mathbb{C}$. Diagonalizable:

$$A = S \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} S^{-1}$$

hence

$$f(A) = S \begin{bmatrix} f(\lambda_1) & 0 & \cdots & 0 \\ 0 & f(\lambda_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & f(\lambda_n) \end{bmatrix}$$

The rotation matrix will have real entries, but complex (unit length) eigenvalues. So it isn't defective.

While with projection onto 2-dimensional plane, you'll have eigenvalue 1 with multiplicitiy 2.

So we have gotten some intuition of eigenvalues. Greater than 1, it is stretching it along the corresponding eigenvector, and less than 1, it is shrinking it. So iterated powers either make it go to infinity, or to 0.

7.1 Recurrence Relations and Finite Precision

7.1.1 Fibonacci numbers:

The Fibonacci numbers are defined as $F_{n+2} = F_{n+1} + F_n$ with the initial conditions $F_1 = F_2 = 1$. So our sequence is $1, 1, 2, 3, 5, 8, 13, 21, 34, \ldots$. We can also extend this backwards for $F_0, F_{-1}, F_{-2}, \ldots$ with the recurrence relation $\ldots, -8, 5, -3, 2, -1, 1, 0, 1, 1, \ldots$; as you can see, backwards follows a similar pattern.

We can use some linear algebra here. Since, we have $F_{n+2} = F_{n+1} + F_n$ and $F_{n+1} = F_{n+1}$, we can write

$$\begin{bmatrix} F_{n+2} \\ F_{n+1} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} F_{n+1} \\ F_n \end{bmatrix}$$

We can iterate backwards until we get to our initial conditions, $\begin{bmatrix} F_2 \\ F_1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} F_2 \\ F_1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, so

$$\begin{aligned} \begin{bmatrix} F_{n+1} \\ F_n \end{bmatrix} &= A \begin{bmatrix} F_n \\ F_{n-1} \end{bmatrix} \\ &= A \cdot A \cdot \begin{bmatrix} F_{n-1} \\ F_{n-2} \end{bmatrix} \\ &\vdots \\ &= A^n \begin{bmatrix} F_1 \\ F_0 \end{bmatrix} \end{aligned}$$

See how this is similar to what we did with Euler's method, except the "dyanmics" dictate that we take the power instead. There is a strong connection to Euler's method, $\vec{y}' = A\vec{y}$. We have

$$\begin{bmatrix} F_{n+1} \\ F_n \end{bmatrix} = A^n \begin{bmatrix} F_1 \\ F_0 \end{bmatrix}$$

How to diagonlize, at 2×2 matrix? Recipe: look for $\vec{v}$ such that there is some λ where $A\vec{v} = \lambda\vec{v} = \lambda \cdot I\vec{v}$. So $(A - \lambda I)\vec{v} = 0$. So we look for λs such that $A - \lambda I$:

(a). has non-zero vector in its nullspace

(b). $\det = 0$

(c). rank $< n$ (for $n \times n$)

(d). not invertible

(e). etc. (your favourite condition)

We can solve (obviously not the move for higher dimensions):

$$\det\left(\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix} - \begin{bmatrix}\lambda & 0\\ 0 & \lambda\end{bmatrix}\right) = 0 \implies \det\begin{bmatrix}1-\lambda & 1\\ 1 & -\lambda\end{bmatrix} = 0$$

So we are solving $(1-\lambda)(-\lambda) - (1)(1) = 0$, or $\lambda^2 - \lambda - 1 = 0$. The solutions to this are the golden ratio and its "conjugate": $\lambda = \frac{1\pm\sqrt{5}}{2}$. So given that $\lambda = \frac{1+\sqrt{5}}{2}$, $\vec{v}$ such that

$$\begin{bmatrix}1-\frac{1+\sqrt{5}}{2} & 1\\ 1 & -\frac{1+\sqrt{5}}{2}\end{bmatrix}\begin{bmatrix}v_1\\ v_2\end{bmatrix} = 0$$

So we want $(1-\frac{1+\sqrt{5}}{2})v_1 + v_2 = 0$. If we take $v_2 = 1$, then $\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}?\\ 1\end{bmatrix} = \frac{1+\sqrt{5}}{2}\begin{bmatrix}?\\ 1\end{bmatrix}$.

Let's look at a simpler recurrence. $F_{n+1} = 10F_n$. Then $F_n = 10^n F_0$. Just like an ODE, if we have an inital condition, reduces to e^{At} where A is a number. In the second order, our A is a matrix.

In our Fibonacci case, we can find $F_n = \left(\frac{1+\sqrt{5}}{2}\right)^n F_0$ and $F_n = \left(\frac{1-\sqrt{5}}{2}\right)^n F_0$. So we have $\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}\lambda\\ 1\end{bmatrix} = \begin{bmatrix}\lambda^2\\ 1\end{bmatrix}$. So one solution is $F_n = c_1\left(\frac{1+\sqrt{5}}{2}\right)^n + c_2\left(\frac{1-\sqrt{5}}{2}\right)^n$.

This is more general than Fibonacci. If

$$\begin{bmatrix}F_1\\ F_1\end{bmatrix} = A\begin{bmatrix}\lambda\\ 1\end{bmatrix} = A\begin{bmatrix}F_1\\ F_0\end{bmatrix}$$

ff (I think using the fact that $F_1 = \lambda F_0$ or something.)

8 Feburary 2

Theorem 4. *Let $Fy' = f(t,y), y(t_0) = y_0$. Then there exists a (unique) solution if f is (Lipschitz) continuous.*

Proof is in Appedix A, but won't go into; it is quite long.

8.1 Recurrences

Theorem 5. *Let $f = f(n,x)$, i.e. $f\colon \mathbb{Z}\times\mathbb{R}\to\mathbb{R}$ and consider $x_{n+1} = f(n,x_n)$ for all $n \geq n_0$ and $x_{n_0} = x_0$. Then, there exists a unique solution $x_{n_0}, x_{n_0+1}, x_{n_0+2}, \dots$, i.e. $x\colon \{n_0, n_0+1, n_0+2, \dots\}\to\mathbb{R}$.*

Proof. Proof is trivial by induction: $x_{n_0+1} = f(n_0, x_{n_0}), x_{n_0+2} = f(n_0+1, x_{n_0+1}), \dots$ are all uniquely determined by f. □

This is one of the places where a existence/uniqueness proof is much easier. Compare this to the proof for ODEs.

Remark 3. By contrast, we can't go $x_{n_0}, x_{n_0+1}, \dots \to x_{n_0-1}$. It is not clear that $x_{n_0} = f(n_0-1, x_{n_0-1})$.

In the case of Euler's method however, this will work.

When we're solving an ODE, can intuitively think about taking little steps, and incremeneting by our f. But it is not clear what the exact relationship between ODE's and recurrences. Perhaps trapezoidal method for an ODE will give different recurrence than Euler's. But it is often that ODEs are the limit (for say Euler's method) of recurrences (depends on $h > 0$). Regardless, there is a lot that happens in common.

Reverse engineer: $(r-3)(r-2) \leftrightarrow x_n = c_1 2^n + c_2 3^n$. Then $r^2 - 5r + 6 = 0$, and so $(\sigma^2 - 5\sigma + 6)(x_n) = 0$, so $x_{n+2} - 5x_{n+1} + 6x_n = 0$.

So saw we want to solve $x_{n+2} - 5x_{n+1} + 6x_n = 0$. Guess (like with ODEs) maybe $x_n = r^n$ will be a solution for some r. Plugging in, we have $r^{n+2} - 5r^{n+1} + 6r^n = 0$. If $r \neq 0$, this holds for all $n \in \mathbb{Z}$ if and only if (???) $r^2 - 5r + 6 = 0 \implies (r-2)(r-3) = 0$. So $x_n = 2^n$ works, 3^n works. Check our work: $x_0 = 1$, $x_1 = 2$, $x_2 = 4$: $4 = 5\cdot 2 - 6\cdot 1$ works ($n = 0$), and $8 = 5\cdot 4 - 6\cdot 2$ works ($n = 1$). Similarly, for $x_n = 3^n$, so $x_n = c_1 2^n + c_2 3^n$.

Proposition 1. *For $x_{n+2} = 5x_{n+1} - 6x_n$ for any x_0, x_1, there is a unique solution.*

Proof. We have x_2 is a function of x_1, x_0, and x_3 is a function of x_2, x_1, and, etc. and $6x_n = 5x_{n+1} - x_{n+2} \implies x_n = (5x_{n+1} - x_{n+2})/6$. And so x_{-1} is a function of x_0, x_1. □

Another:

Proof. We know $x_n = c_1 2^n + c_2 3^n$ is a solution. So find c_1, c_2 that works:

$$c_1 2^0 + c_2 3^0 = x_0$$
$$c_1 2^1 + c_2 3^1 = x_1$$

or x_0, x_1 given, solve

$$\begin{bmatrix} 1 & 1 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix}$$

This is invertible, so solvable. Even more nice, this is a Vandermonde matrix! Will show up everywhere. □

Similarly, if we had $x_n = c_1 2^n + c_2 3^n + c_3 7^n$ via $(\sigma - 2)(\sigma - 3)(\sigma - 7)(x_n) = 0$ to solve for c_1, c_2, c_3 given x_0, x_1, x_2,

$$\begin{bmatrix} 1 & 1 & 1 \\ 2 & 3 & 7 \\ 2^2 & 3^2 & 7^2 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix}$$

And get $x_{n+3} = \cdots x_{n+2} \cdots x_{n+1} \cdots x_n$.

Let's return to our recurrence $x_{n+2} = 5x_{n+1} - 6x_n$ or $x_{n+2} - 5x_{n+1} + 6x_n = 0$. So let $\vec{z}_n = \begin{bmatrix} x_{n+1} \\ x_n \end{bmatrix}$, $n \in \mathbb{Z}$.

$$\vec{z}_{n+1} = \begin{bmatrix} x_{n+1} \\ x_n \end{bmatrix} = \begin{bmatrix} 5 & -6 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_{n+1} \\ x_n \end{bmatrix} = \vec{z}_n \begin{bmatrix} x_{n+1} \\ x_n \end{bmatrix}$$

Where we found the entries of our matrix from the relatoin above. We guessed $x_n = r^n$ for $(r = 2, 3)$ is a solution

$$\vec{z}_n = \begin{bmatrix} x_{n+1} \\ x_n \end{bmatrix} = \begin{bmatrix} r^{n+1} \\ r^n \end{bmatrix}$$
$$\vec{z}_{n+1} = \begin{bmatrix} 5 & -6 \\ 1 & 0 \end{bmatrix} \vec{z}_n$$
$$\begin{bmatrix} r^{n+2} \\ r^{n+1} \end{bmatrix} = \begin{bmatrix} 5 & -6 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} r^{n+1} \\ r^n \end{bmatrix}$$

Which is true if and only if $(r \neq 0)$

$$\begin{bmatrix} r \\ 1 \end{bmatrix} = \begin{bmatrix} 5 & -6 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} r \\ 1 \end{bmatrix}$$

which is to say that $\begin{bmatrix} r \\ 1 \end{bmatrix}$ is an eigenvector. So if $r = 2, 3$ work:

$$2 \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 5 & -6 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

$$3 \begin{bmatrix} 3 \\ 1 \end{bmatrix} = \begin{bmatrix} 5 & -6 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

Computing the eigenvalues of $\begin{bmatrix} 5 & -6 \\ 1 & 0 \end{bmatrix}$: $\det(\lambda I - \begin{bmatrix} 5 & -6 \\ 1 & 0 \end{bmatrix})$. Which gives $0 = \lambda^2 - 5\lambda_6$.

Does this method always work? Consider $x_{n+2} = 4x_{n+1} - 4x_n$, so $x_{n+2} - 4x_{n+1} + 4x_n = 0$. We see then $0 = (\sigma^2 - 4\sigma + 4)(x_n) = (\sigma - 2)^2(x_n)$. Duplicate root. We guess $x_n = 2^n$: this will work, but what else will? There's no r^n for $r \neq 2$ $r^2 - 4r + 4 = (r - 2)^2 = 0$. What else might work? We will answer this on Monday.

9 February 5

3 Handouts:

- Intro to ODEs
- Recurrence Relations and Finite precision (you'll notice no mention of ODEs; from old course order)... some of the next homework problems are on here, and some have partial solutions too. We did Appendices A, B, and C
- Normal and Subnormal numbers

Today: A bit more on first two handouts, and start normal and subnormal numbers.

3-term recurrence relations: $x_{n+2} = ax_{n+1} + bx_n$, $a, b \in \mathbb{R}$, $\ldots, x_{-1}, x_0, x_1, \ldots$, $b \neq 0$ (then if $a \neq 0$, $x_{n+2} = ax_{n+1}$ which is just one term). This is actually an analog of second order ODE: $\vec{y}_{n+1} = A\vec{y}_n$ where $\vec{y}_{n+1} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} F_{n+1} \\ F_n \end{bmatrix}$. ff something I was busy on Piazza

Last time: $x_{n+2} - 4x_{n+1} + 4x_n = 0$, $(\sigma^2 - 4\sigma + 4)(x_n) = 0$. Guess $x_n = r^n$ is a solution... $r^2 - 4r + 4 = 0$, $(r-2)^2 = 0$, $r = 2, 2$. So $x_n = 2^n$ is a solution, and $x_n = c_1 2^n$ is also a solution. Think of it... $(r-2)(r-2-\varepsilon) = 0$. If $\varepsilon > 0$: $(\sigma - 2)(\sigma - 2 - \varepsilon) = 0$, then $x_n = c_1 2^n + c_2(2+\varepsilon)^n$. Fix $\varepsilon > 0$, have $x_0 = 3, x_1 1 = -40, x_2 = x_2(\varepsilon), x_3, \ldots$. So $x_{21} = (\cdot)x_0 + (\cdot)x_1$ given x_0, x_1. Taking $\varepsilon \to 0$, $x_{21}(0) = (\cdot)x_0 + (\cdot)x_1$.

What is the general solution to $x_{n+2} - 4x_{n+1} + 4x_n = 0$ if $x_n = c_1 2^n + c_2 n 2^n$ (like what we do with ODEs). Does this work? Well, $(n+2)2^n - 4(n+1)2^{n+1} - 4 \cdot 2^n = 0$. Try this: $(\sigma - 2)(\sigma - 2)(n2^n) = 0$? $(\sigma - 2)(n2^n) = (n+1)2^{n+1} \cdots 2(n2^n) = 2^n((n+1)2 - 2n) = 2^n(2)$ and $(\sigma - 2)(2^n \cdot \text{constant}) = \text{constant}(2^{n+1} - 2 \cdot 2^n) = 0$. What about $(\sigma - 2)^4(x_n) = 0$. $x_n = c_1 2^n + c_2 n 2^n + c_3 n^2 2^n + c_4 n^3 2^n 4$.

In general, we have for $p(n)$ a polynomial, $(\sigma - 2)(p(n)2^n) = 2^n \cdot$(polynomial of lower degree). We can verify for our case $(\sigma - 2)(n^k 2^n) = (n+1)^k 2^{n+1} - n^k 2^n \cdot 2 = 2^{n+1}((n+1)^k - n^k) = 2^{n+1}(kn^{k-1} + \text{lower})$. Hmm... kx^{k-1} looks like the derivative of x^k. When we interpolate a polynomial, take two nearby points and find the value in between, we are taking the derivative.

Now, the general solution to $(\sigma - r_1)^{m_1} \cdots (\sigma - r_k)^{m_k}(x_n) = 0$ is

$$x_n = p_1(n)r_1^n + \cdots + p_k(n)r_k^n$$

where $\deg(p_i) \leq m_k - 1$. E.g. $(\sigma - 2)^2(\sigma - 5)(x_n) = 0$ has the solution $x_n = c_1 2^n + c_2 n 2^n + c_3 5^n$.

Looking at $x_{n+2} - 4x_{n+1} + 4x_n = 0$, $\vec{y}_n = \begin{bmatrix} x_{n+1} \\ x_n \end{bmatrix}$, $\vec{y}_{n+1} = \begin{bmatrix} x_{n+2} \\ x_{n+1} \end{bmatrix} = \begin{bmatrix} 4 & -4 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_{n+1} \\ x_n \end{bmatrix}$. ff

$\lambda^2 - 4\lambda + 4 = 0$, $\lambda = 2, 2$. We know then that we cannot have two linearly independent eigenvectors. When we have two eignevalues that are the same, cannot be linearly independent, otherwise $A = S \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} S^{-1} = 2SIS^{-1} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$. (Why can't we do this?)

Look at $\lambda I - \begin{bmatrix} 4 & -4 \\ 1 & 0 \end{bmatrix}$. ff We call these matrices deficient, defective, etc. We cannot diagonalize this matrix.

Since $2I - \begin{bmatrix} 4 & -4 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} -2 & 4 \\ -1 & 2 \end{bmatrix} =: N$. We have $N \neq \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$, but $N^2 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$. We say that N is *nilpotent* if $N^k = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$ for some $k \geq 1$.

So we have $2I - \big[\text{matrix of interest}\big] = N = \text{nilpotent}$. ff

10 February 16

Not Joel, the 302 guy, but he's one of the author's of the textbook.

Recall monomial interpolation: we are given two points, can easily find $c_0 + c_1 x$ such that it passes through both. If $(1,1), (2,3)$, then we let $1 = c_0 + c_1 \cdot 1$ and $3 = c_0 + c_1 \cdot 2$, and solve the linear system $\begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$. Then $p_1(x) = 2x - 1$. Similar happens if there is a third point and we use a degree 2 polynomial: $p_2(x) = c_0 + c_1 x + c_2 x^2$.

These are super-intuitive: if asked to find a polynomial, this is what you would do. But this method is not always the best. Some problems:

(1) Can't directly connect solution c_0, c_1 to the data points, not obvious what the connection is (this will be clearer when we talk about Lagrange points later today).

(2) But also, for an $n \times n$ system, our cost is $O(n^3)$ flops (floating point operations?). This is not that good, but not awful because our computers can evaluate a polynomial really fast. But it is high cost, because even though n is typically not that large, we often need to repeat many times and it becomes expensive (e.g. in machine learning, computing many polynomials). However, although the construction is expensive, the evaluation is cheap (can easily evaluate the polynomial at some other value).

(3) In MATLAB, we have the command `vander` to get the vandermonde matrix of ??? and `cond` to get that matrix's condition number. Now can compute something: the matrix is badly condition (which means inaccurate computations). Something like close to 10^{16} (particularly bad).

Recall the definition of the condition number K: $K(A) = \|A\|\|A^{-1}\|$. Suppose $Ax = b$, $\tilde{x}$ is a numerical solution. Then

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leq K(A) \frac{\|b - A\tilde{x}\|}{\|b\|}$$

In our previous example, x is our solution $\begin{bmatrix} c_0 \\ c_1 \end{bmatrix}$. In a common application, we can't compute x, and have $\tilde{x}$. And we can compute the residual $A\tilde{x}$. So we can compute $\frac{\|b-A\tilde{x}\|}{\|b\|}$. But if $K(A)$, we do not have any assurance that our $\tilde{x}$ is a good approximation of x.

Some facts about condition numbers: the smallest they can be is $K(A) = 1$, called perfectly conditioned. See $\|I\| = 1$. The worst possible is $K(A) = \infty$, this is for singular matrices.

10.1 Lagrange interpolation

Quite a nice method, very different from monomial. We are given the points $(x_j, y_j)_{j=0}^n$. Our monomial interpolation gave us

$$p_n(x) = \sum_{j=0}^{n} c_j \phi_j(x)$$

where $\phi_j(x) = x^{j-1}$. Lagrange interpolation gives

$$p_n(x) = \sum_{j=0}^{n} y_j L_j(x)$$

Notably now, we actually get the y_js showing up in the solution (remember the relation between the polynomial and the points?). How do we construct our L_js? We want

$$L_j(x_i) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

Then $p_n(x_3) = y_3$, etc. We want to construct Lagrange polynomials that do this for us.

What is L_0 for our system from before $((1,1),(2,3),(4,3))$? We can make 2 and 3 roots, and have a coefficient out front to normalize. So $L_0(x) = a(x-2)(x-4)$, and $L_0(1) = 1 = a(1-2)(1-4) \implies a = \frac{1}{3}$. So in the arbitrary setting, we have

$$L_j(x) = \frac{1}{(x_j - x_0)\cdots(x_j - x_n)}(x - x_0)\cdots(x - x_{j-1}(x - x_{j+1})(x - x_n)$$

What is the linear system we solved? It was the identity matrix. What was the cost of construction? Well, we did n multiplications n times, and so there were $O(n^2)$ floating point operations. But the evaluation is more costly: now need to evaluate n^2.

Something about uniqueness, but didn't get to.

11 February 26

Plan:

- Review:
  - Relative error in double precision
  - Relative error and condition number

$$K_p(A) := \|A\|_p \|A^{-1}\|_p$$

- Compare
  - 10.2 Monomial Interpolation
  - 10.3 Lagrange Interpolation
- This week: 10.4 Divided Differences

Here is the basic problem of this course: CPSC 303 is not supposed to be 302. But things like condition numbers are so useful, that we want to use them, and will mainly see their application towards interpolation.

Chen is more like a numerical algebraist. When he sees condition numbers, particularly an example of a bad condition number, he gets excited. Joel really just sees it as a tool that something is going wrong.

Remark 4. Say you have scientific notation, base 10, remembering 4 digits: $1.00 \times 10^{71}, 1.001 \times 10^{71}, \ldots, 9.999 \times 10^{71}, 1.000 \times 10^{72}$. So for something with true value 3.217569×10^{71}, the scienticific notation up to 4 places is 3.218×10^{71}, which gives the relative error

$$\frac{|(\text{true value}) - (\text{sci notation})|}{|\text{true value}|} = \frac{|3.217569 - 3.218|}{|3.217569|} \leq \frac{|0.0005|}{3.217569|}$$

So the maximum size of the relative error is

$$\frac{|\frac{1}{2} \cdot 10^{-3}|}{|1.000|} = \frac{1}{2} \cdot 10^{-3}$$

Similar to in the remark, the error in double precision for a standard number:

$$\pm 1.b_1 \ldots b_{52} \cdot 2^m \quad (-1022 \leq m \leq 1023)$$

has maximum relative error

$$2^{-52} \cdot \frac{1}{2} = 2^{-53} \approx 1.1 \times 10^{-16}$$

In ODE's and recurrences, this error compounds.

Let's say we are solving $A\vec{x}_{\text{true}} = \vec{b}_{true}$, but really $\vec{b}_{\text{true}} \rightsquigarrow \vec{b}_{\text{observed}} = \vec{b}_{\text{true}} + \vec{b}_{\text{error}}$. So what we are solving is $A\vec{x}_{\text{observed}} = \vec{b}_{\text{observed}} \neq \vec{b}_{\text{true}}$ (we have the true value of A though)... I think this is a definition for $\vec{x}_{\text{observed}}$. Then $\vec{x}_{\text{observed}} = A^{-1}\vec{b}_{\text{observed}}$. The condition number answers how bad does this relative error of $\vec{x}$ differ. Specifically: $1 \leq p \leq \infty$, using $\|\cdot\|_p$. The relative error in $\vec{b}$:

$$\text{Relative error in } \vec{b} = \frac{\|\vec{b}_{\text{true}} - \vec{b}_{\text{observed}}\|}{\|\vec{b}_{\text{true}}\|}$$

Then

$$\frac{\|\vec{x}_{\text{true}} - \vec{x}_{\text{observed}}\|}{\|\vec{x}_{\text{true}}\|} \leq C \frac{\|\vec{b}_{\text{true}} - \vec{b}_{\text{observed}}\|}{\|\vec{b}_{\text{true}}\|}$$

And we call the maximum C the condition number. We might ask, what is the max C, and when is this max attained?

Last week: $1 \leq p \leq \infty$, we said the worst C is

$$C = K_p(A) = \|A\|_p \|A^{-1}\|_p$$

Let's just believe this... Something something decoupled??? We can solve the error and true value seperately or something.

We will actually see in the homework that the condition number for a Vandermonde matrix is really bad. The homework gives a matrix where we can compute it, at least for one row (bottom righ entry???), which will help show how bad things can actually get. The monomial method relies on Vandermonde matrices, and so get a high error. We will see the Lagrange and divided differences have a much lower condition number.

Trick we will use: Recall $A\vec{x}_{\text{true}} = \vec{b}_{\text{true}}$ and $A\vec{x}_{\text{observed}} = \vec{b}_{\text{observed}}$. And so $A(\vec{x}_{\text{true}} - \vec{x}_{\text{obs}}) = \vec{b}_{\text{true}} - \vec{b}_{\text{obs}}$, and if we define $\vec{b}_{\text{error}} := \vec{b}_{\text{true}} - \vec{b}\text{obs}$, then $A\vec{x}_{\text{error}} = \vec{b}_{\text{error}}$.

So now we can express our problem in terms of the relative error in $\vec{x}$ with the relative error in $\vec{b}$ (looking for max C):

$$\frac{\|\vec{x}_{\text{error}}\|_p}{\|\vec{x}_{\text{true}}\|_p} \leq C \frac{\|\vec{b}_{\text{error}}\|_p}{\|\vec{b}_{\text{true}}\|_p}$$

We can now seperate our problem:

(1) What is the max constant C_1 such that

$$\|\vec{x}_{\text{error}}\|_p \leq C_1 \|\vec{b}_{\text{error}}\|_p$$

(2) What is the max constant C_2 such that

$$\frac{1}{\|\vec{x}_{\text{true}}\|_p} \leq C_2 \frac{1}{\|\vec{b}_{\text{true}}\|_p}$$

Where $\vec{b}_{\text{error}}$ is anything, and $\vec{b}_{\text{true}}$ is anything, so these are completely independent of each other! Probably the most remarkable thing in this course. On the homework, they will give a few 2×2 examples where this happens.

Looking at the error first, we have $\vec{x}_{\text{error}} = A^{-1}\vec{b}_{\text{error}}$, so

$$\|\vec{x}_{\text{error}}\|_p = \|A^{-1}\vec{b}_{\text{error}}\|_p \leq \|A^{-1}\|_p \|\vec{b}_{\text{error}}\|_p$$

Also, you can have $\vec{x}_{\text{error}}, \vec{b}_{\text{error}}$ non-zero, and

$$\|A^{-1}\vec{b}_{\text{error}}\|_p = \|A^{-1}\|_p \|\vec{b}_{\text{error}}\|_p$$

since

$$\|M\|_p := \max_{\vec{x} \neq 0} \frac{\|M\vec{x}\|_p}{\|\vec{x}\|_p}$$

Remark 5. To find such a pair $\vec{b}_{\text{error}}, \vec{x}_{\text{error}} = A^{-1}\vec{b}_{\text{error}}$, it is not so nice for $\|\cdot\|_2$, but not so bad for $\|\cdot\|_\infty$, since

$$\left\| \begin{bmatrix} a & b \\ c & d \end{bmatrix} \right\|_\infty = \max(|a| + |b|, |c| + |d|)$$

So $\|\vec{x}\|_p \leq C_1 \|\vec{b}_{\text{error}}\|_p$ where C_1 can be as low as $\|A^{-1}\|_p$.

Now, we want the smallest possible C_2 such that

$$\frac{1}{\|\vec{x}_{\text{error}}\|_p} \leq C_2 \frac{1}{\|\vec{b}_{\text{error}}\|_p}$$

I.e. the smallest C_2 such that

$$\|\vec{b}_{\text{true}}\|_p \leq C_2 \|\vec{x}_{\text{true}}\|_p$$

But recall $\|\vec{b}_{\text{true}}\|_p$ ff So smallest C_2 is $\|A\|_p$, and $\|\vec{b}_{\text{true}}\|_P \leq \|A\|_p \|\vec{b}_{\text{true}}\|_P$ is attained with equality when

$$\|A\vec{x}_{\text{true}} = \|A\|_p \|\vec{x}_{\text{true}}\|_p$$

(it is stretching $\vec{x}_{\text{true}}$ out to its maximum).

So we have $C_1 = \|A^{-1}\|_p$ and $C_2 = \|A\|_p$. ff

12 February 28

For today:

- Monic interpolation (Section 10.2) versus Lagrange interpolation (Section 10.3)
- Divided differences (Sections 10.4 - 10.7) (See article: CPSC 303 Remarks on divided differences. Likely to be revised to include more comments on 10.4-10.7, since book doesn't even talk about it enough)

Homework 7 (not yet posted), will likely look at the following...

Monic interpolation: given data $(x_0, y_0), \dots, (x_n, y_n)$ and want to fit a polynomial $p(x)$ of degree $n+1$ (should be n???), specifically $p(x) = c_0 + c_1 x + c_2 x^2 + \cdots + c_n x^n$, such that $p(x_i) = y_i$, $i = 0, \dots, n$. In 10.2, we use the Vandermonde matrix

$$\begin{bmatrix} 1 & x_0 & \cdots & x_0^n \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \cdots & x_n^n \end{bmatrix} \begin{bmatrix} c_0 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} y_0 \\ \vdots \\ y_n \end{bmatrix}$$

What could possible go wrong? Say $(x_0, y_0), (x_1, y_1)$, and $x_0 = 2, x_1 = 2 + 10^{-5}$. You just want $c_0 + c_1 x$, so to find c_0, c_1 compute halfway between: $p(2 + (10^{-5})\frac{1}{2})$, call this point x_{mid}. Our system is (??? why did he put this here)

$$\begin{bmatrix} 1 & 2 \\ 1 & 2 + 10^{-5} \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \end{bmatrix}$$

but we can compute $p(x_{\text{mid}}) = \frac{y_0 + y_1}{2}$ since it is linear.

We can numerically solve our system: $A\vec{c} = \vec{y}$, where A is our matrix from above. We can solve $A^{-1} = \frac{1}{\det} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \frac{1}{(2+10^{-5})-2} \begin{bmatrix} 2+10^{-5} & -2 \\ -1 & 1 \end{bmatrix} = 10^5 \begin{bmatrix} 2+10^{-5} & -2 \\ -1 & 1 \end{bmatrix}$. It is clear to see that this matrix will have a very large condition number. We can compute the infinty norm (since it is easiest) to see

$$\left\| \begin{bmatrix} 1 & 2 \\ 1 & 2 + 10^{-5} \end{bmatrix} \right\|_\infty = \max(1 + 2, 1 + 2 + 10^{-5}) = 3 + 10^{-5}$$

and so ff some problem

But consider Lagrange interpolation: say fitting $(x_0, y_0), (x_1, y_1), (x_2, y_2)$. Note fixed x_0, x_1, x_2. We have

$$L_0(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)}$$

satisfies $L_0(x_1) = (x_1 - x_1)(\text{etc.}) = 0$, $L_0(x_2) = 0$, and $L_0(x_0) = \frac{(x_0 - x_1)(x_0 - x_2)}{(x_0 - x_1)(x_0 - x_2)} = 1$ (note that this is still a polynomial, since the denominator is just a constant). So this is like the Dirac delta. Then, we have

$$p(x) = y_0 L_0(x) + y_1 L_1(x) + y_2 L_2(x)$$

Then, it is easy to compute that $p(x_0) = y_0, p(x_1) = y_1, p(x_2) = y_2$. Note that the L_i are all polynomials of degree 2.

We will see that Lagrange interpolation solves our problem. Let's look at the linear case $(x_0, y_0), (x_1, y_1)$. Then $p(x) = y_0 \frac{x - x_1}{x_0 - x_1} + y_1 \frac{x - x_0}{x_1 - x_0}$. This is the exact same polynomial as before with monic interpolation, and will be the same one we get with divided differences. What happens: $x_0 = 2, x_1 = 2 + 10^{-5}$, and want $p(x_{\text{mid}}) = p(2 + 10^{-5}\frac{1}{2})$. We compute it the way Lagrange suggests to do: compute each term seperately first:

$$p(x_{\text{mid}}) = y_0 \left(\frac{x_{\text{mid}} - x_1}{x_0 - x_1} \right) + y_2 \left(\frac{x_{\text{mid}} - x_0}{x_1 - x_0} \right)$$

The first term becomes $\frac{((2+10^{-5}\frac{1}{2}) - (2+10^{-5}))}{(2-(2+10^{-5}))}$. MATLAB, using double precision, has maximum relative error 10^{-16}. We can calculat $2 + 10^{-5}\frac{1}{2} = 2.000005$. then we are computing the difference $2.000005 - 2.00001$, each with at least 10^{-16} relative precision, so get $-0.000005 \pm 2 \cdot 10^{-16}$.

The point:

$$\frac{(x-x_0)(x-x_1)\cdots(x-x_{n-1})}{(x_n-x_0)(x_n-x_1)(\cdots(x_n-x_{n-1}}$$

and $x_0, \ldots, x_n$ close: $x_0 = 2, x_1 = 2+\varepsilon, x_2 = 2+2\varepsilon, \ldots$. And $x \to 2+\varepsilon/2$???

Even though the polynomials are mathematically the same, Lagrange interpolation is computed much better. Perhaps if you wanted to evaluate a ton of points, and they're not all close together, monic interpolation might be better. But this gives one siutation where Lagrange interpolatoin loses less precision than monic interpolation.

On the homework, will choose y_0, y_1 to be some irrational number, so that precision is forced to be lost. Some fractions in base 2 are exact (like $1/8$), but all irrationals cannot be exact in any base.

Moving on, 10.4 - 10.7 and beyond is a much bigger story of Newton's divided differences.