CPSC 303: INTRODUCTION TO ODE'S (AND REVIEW OF CALCULUS FOR NUMERICAL METHODS)

JOEL FRIEDMAN

Contents

Date: Monday 8th April, 2024, at 17:46(get rid of time in final version).
Research supported in part by an NSERC grant.

Disclaimer: The material may sketchy and/or contain errors, which I will elaborate upon and/or correct in class. For those not in CPSC 303: use this material at your own risk...

The goal of this article is to review some basic notions from calculus that we will use in CPCS 303, and to present part of the material that is representative of the level of difficulty of this course.

[The usual first few weeks of CPSC 303 tend to be misleadingly easy.]

One of the two main goals of CPSC 303 is the study of differential equations, and much of the emphasis will be on ODE's (Ordinary Differential Equations). In this article we will give an introduction to ODE's, and mention PDE's.

The other main goal is to study curve fitting with polynomials or piecewise polynomials, which involves Taylor's theorem and—at times—some "variational calculus." Hence we review the ideas we will need.

Throughout this article, [A&G] refers to the course textbook by Ascher and Greif. We tend to use the same notation in [A&G].

The UBC Mathematics Department has calculus textbooks publicly available at: https://personal.math.ubc.ca/~CLP/.

1. Preliminary Ideas from [A&G]

Most of the preliminary ideas we discuss here are based on Section 1.2 of [A&G]. We also briefly mention ℓ^p-norms of Section 4.2 and numerical differentiation of Section 14.1. We will also introduce some common notation.

All the above will be used throughout this course.

1.1. Relative and Absolute Error.

If $u \in \mathbb{R}$ is approximated by $v \in \mathbb{R}$, then the *absolute error in v (as an approximation of u)* is $|u-v|$, and the relative error is $|u-v|/|u|$.

The same definition holds for $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ and the *ℓ^p-norm* on $\mathbb{R}^n$. (The textbook [A&G] uses the notation ℓ_p-norm, but ℓ^p is far more common in the literature.) If $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$, Section 4.2, page 74 of [A&G] defines

$$\|\mathbf{x}\|_1 \stackrel{\text{def}}{=} |x_1| + \cdots + |x_n|,$$
$$\|\mathbf{x}\|_2 \stackrel{\text{def}}{=} \sqrt{x_1^2 + \cdots + x_n^2},$$
$$\|\mathbf{x}\|_\infty \stackrel{\text{def}}{=} \max_{1 \le i \le n} |x_i|.$$

More generally, for any $1 \le p < \infty$ one defines the ℓ^p-norm (sometimes ℓ_p-norm) of $\mathbf{x} \in \mathbb{R}^n$ as

$$\|\mathbf{x}\|_p \stackrel{\text{def}}{=} \Big(|x_1|^p + \cdots + |x_n|^p \Big)^{1/p},$$

which generalizes the ℓ^1-norm and ℓ^2 norm above, and one also sees that for any $\mathbf{x}$, as $p \to \infty$ we have $\|\mathbf{x}\|_p \to \|\mathbf{x}\|_\infty$.

Hence if $u \in \mathbb{R}^n$ is approximated by $v \in \mathbb{R}^n$, and $1 \leq p \leq \infty$, then the *absolute ℓ^p-error in v (as an approximation of u* is $\|u - v\|_p$, and the relative error ℓ^p is $\|u - v\|_p / \|u\|_p$.

The ℓ^p-norm will be needed when we discuss *condition numbers* of matrices. Note that for $x \in \mathbb{R} = \mathbb{R}^n$ with $n = 1$, $\|x\|_p = |x|$ for any p; hence the ℓ^p-norm generalizes the usual absolute value.

1.2. **Taylor's Theorem.** In Chapter 1 of [A&G], page 5, Taylor's Theorem is written as: follows: assume $k \in \mathbb{N} = \{1, 2, \ldots\}$, $f\colon (a, b) \to \mathbb{R}$ has $k+1$ derivatives, and $x_0, h \in \mathbb{R}$ with $x_0, x_0 + h \in (a, b)$; then there exists a ξ between x_0 and $x_0 + h$ such that

$$f(x_0 + h) = f(x_0) + hf'(x_0) + \frac{h^2}{2} f''(x_0) + \cdots + \frac{h^k}{k!} f^{(k)}(x_0) + \frac{h^{k+1}}{(k+1)!} f^{(k+1)}(\xi)$$

(moreover, if $h \neq 0$ then one can find a ξ strictly bewteen x_0 and $x_0 + h$).

A consequence of this theorem is *Taylor series*: say that for f, a, b, x_0, h as above, and all $k \in \mathbb{N}$, there is an $M \in \mathbb{R}$ such that $|f^{(k)}(\xi) \leq k! M^k$. Then for each $h \in \mathbb{R}$ with $h < 1/M$ and all $x_0 \in \mathbb{R}$ with $x_0, x_0 + h \in (a, b)$, we have

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

which is called the *Taylor series* of f at x_0.

Example 1.1. For example, let $f(x) = e^x$. Then we have

$$f'(x) = f''(x) = \ldots = e^x,$$

and hence for $x_0 = 0$, for any h we have

$$\begin{aligned} f(0 + h) &= f(0) + hf'(0) + \frac{h^2}{2} f''(0) + \cdots + \frac{h^k}{k!} f^{(k)}(0) + \frac{h^{k+1}}{(k+1)!} f^{(k+1)} \xi \\ &= 1 + h + \frac{h^2}{2} + \cdots + \frac{h^k}{k!} + \frac{h^{k+1}}{(k+1)!} e^{\xi}, \end{aligned}$$

for some ξ between $x_0 = 0$ and $x_0 + h = h$. So for h fixed, $e^{\xi} \leq e^{|h|}$ (note that h can be negative), and

$$\frac{h^{k+1}}{(k+1)!} e^{\xi} \leq e^{|h|} \frac{h^{k+1}}{(k+1)!} = e^{|h|} \frac{h}{1} \frac{h}{2} \cdots \frac{h}{k+1}.$$

In particular, if $h \neq 0$ and we fix a $K \geq 2/h$, say $K = \lceil 2/h \rceil$ (the "ceiling function"), we have $h/K \leq 1/2$, and therefore

$$e^{|h|} \frac{h}{1} \frac{h}{2} \cdots \frac{h}{k+1} = \left(e^{|h|} \frac{h}{1} \cdots \frac{h}{K-1} \right) \frac{h}{K} \cdots \frac{h}{k+1} \leq \left(e^{|h|} \frac{h}{1} \cdots \frac{h}{K-1} \right) (1/2)^{k+1-K+1}$$

which tends to 0 as $k \to \infty$ for any $h \neq 0$. Hence

$$e^h = 1 + h + \frac{h^2}{2} + \frac{h^3}{3!} + \ldots$$

is valid for all h. Since it is valid for any $h \in \mathbb{R}$, we tend to write this equation as

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \ldots$$

(since in this course h is usually a small number, and x is usually a real variable).

Remark 1.2. Another way to see that the above Taylor series converges is to use Stirling's approximation that $n! \sim (n/e)^n \sqrt{2\pi n}$[1] Hence, for any f and x_0, the Taylor expansion for $f(x_0+h)$ at $x=x_0$ converges for all h with $|h| \le h_0$ for any h_0 such that

$$\frac{h_0^k}{(k/e)^k\sqrt{2\pi k}} \max_{|\xi - x_0| \le h_0} \left| f^{(k)}(\xi) \right|$$

tends to 0 as $k \to \infty$.

Similarly, we have

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \dots, \qquad \cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \dots.$$

Both of these can also be derived from the expansion for e^x, using the fact that if $z \in \mathbb{C}$, then one has the well-known equality in complex variables

$$e^{iz} = \cos(z) + i\sin(z).$$

Remark 1.3. The function $\mathbb{R} \to \mathbb{R}$ given by

$$f(x) = \begin{cases} e^{-1/x^2} & \text{if } x \ne 0, \text{ and} \\ 0 & \text{if } x = 0 \end{cases}$$

is infinitly differentiable and satisfies $f(0) = f'(0) = f''(0) = \dots = 0$. However $f(h) \ne 0$ for $h \ne 0$. Hence this f does not have a Taylor series at $x_0 = 0$.

2. First Derivative Approximation

Section 14.1 of [A&G] uses Taylor's theorem to prove two basic ways to approximate the derivative.

A consequence of Taylor's theorem is that if $f\colon (a,b) \to \mathbb{R}$ is a twice differentiable function, then for any $x_0, h \in \mathbb{R}$ with $x_0, x_0+h \in (a,b)$ and $h \ne 0$ we have

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

for some ξ strictly between x_0 and x_0+h. Hence

$$\frac{f(x_0+h)-f(x_0)}{h} = f'(x_0) + \frac{h}{2} f''(\xi).$$

We also write the above as

$$\frac{f(x_0+h)-f(x_0)}{h} = f'(x_0) + O(h),$$

where $O(h)$ means "order h," which is explained in Section 1.2, page 7 of [A&G]; more formally, $O(h)$ denotes any function of h and other parameters that is bounded by $C|h|$ for a constant C (independent of h and all other variable parameters), for h sufficiently small.

In more detail, say that for some $M_2 \in \mathbb{R}$ we have $|f''(x)| \le M_2$ for all $x \in (a,b)$; then

$$\left| \frac{f(x_0+h)-f(x_0)}{h} - f'(x_0) \right| \le (M_2/2)h$$

for $|h|$ sufficiently near 0.

[1] We write $f(n) \sim g(n)$, to mean that as n tends to ∞, $f(n)/g(n)$ tends to 1.

This should not be surprising; indeed, even if f has only one derivative in (a, b), then

$$f'(x_0) = \lim_{h\to 0} \frac{f(x_0+h) - f(x_0)}{h}.$$

Note that numerically, things are a bit different, due to finite precision (this will be covered in greater detail later). Indeed, Section 1.2 of [A&G], pages 7 and 8, warns you that although

$$g(h) \stackrel{\text{def}}{=} \left|\cos(1.2) - \frac{\sin(1.2+h) - \sin(1.2)}{h}\right| \xrightarrow{h\to 0} 0,$$

the MATLAB computed value of $g(h)$ for $h = 10^{-n}$ with $n \in \mathbb{N}$ decreases in n for $n \leq 8$ but increases for $n \geq 8$.

Note, however that if f is three times differentiable, then

$$\frac{f(x_0+h) - f(x_0-h)}{2h} = f'(x_0) + \frac{h^2}{6} f'''(\xi)$$

for some ξ between $x_0 \pm h$ (see [A&G], Subsection 14.1.2 ("Three Point Formulas", centred formula), page 411. Hence

$$\left|\frac{f(x_0+h) - f(x_0-h)}{2h} - f'(x_0)\right| \leq (M_3/6)h^2,$$

where M_3 is a bound on $|f'''|$ in the interval $x_0 \pm h$. More succinctly we may write

$$f'(x_0) = \frac{f(x_0+h) - f(x_0-h)}{2h} \pm \frac{M_3 h^2}{6}.$$

Similarly, we sometimes need the non-centred formula (same page 411, [A&G])

$$f'(x_0) = \frac{-3f(x_0) + 4f(x_0+h) - f(x_0+2h)}{2h} \pm \frac{M_3 h^2}{3},$$

where M_3 is a bound on $|f'''|$ in the interval between x_0 and $x_0 + 2h$.

Remark 2.1. A bound of the form $O(M_3 h^2)$ may seem better than one of the form $O(M_2 h)$. However, if $|f'''|$ is large or does not exist, a bound, M_2, on $|f''|$ may be better in practice. [We may WRITE A HOMEWORK PROBLEM regarding this.]

2.1. **Designing Derivative Schemes, Via Linear Algebra.** Say that we wish to use the values of $f(x_0), f(x_0 \pm h)$ to design a derivative scheme. Then we can convert this into a linear algebra problem. Namely we write:

$$\begin{aligned}
f(x_0) &= f(x_0) \\
f(x_0+h) &= f(x_0) + hf'(x_0) + \frac{h^2}{2} f''(x_0) + O(h^3)M_3 \\
f(x_0-h) &= f(x_0) - hf'(x_0) + \frac{h^2}{2} f''(x_0) + O(h^3)M_3
\end{aligned}$$

and hence for any constants $c_{-1}, c_0, c_1 \in \mathbb{R}$ we have

$$\begin{aligned}
c_{-1}f(x_0-h) + c_0 f(x_0) + c_1 f(x_0+h) = f(x_0)&\ \left(c_{-1} + c_0 + c_1\right) \\
+hf'(x_0)&\ \left(-c_{-1} + c_1\right) \\
+(h^2/2)f''(x_0)&\ \left(c_{-1} + c_1\right) + O(h^3)
\end{aligned}$$

(where the $O(h^3)$ depends on c_1, c_0, c_{-1} and a bound (M_3) on $f'''(x)$ in the relevant interval). So if we want a formula for $hf'(x_0)$ that is valid to within $O(h^3)$, we solve the system

$$c_{-1} + c_0 + c_1 = 0, \quad -c_{-1} + c_1 = 1, \quad c_{-1} + c_1 = 0.$$

This system has a unique solution $c_1 = 1/2$, $c_0 = 0$, and $c_{-1} = -1/2$, which, upon dividing the system above by h yields

$$\frac{(-1/2)f(x_0 - h) + (0)f(x_0) + (1/2)f(x_0 + h)}{h} = f'(x_0) + O(h^2),$$

which is the *centred "three-point" formula* at the top of page 411, Subsection 14.1.2, of [A&G] (we don't actually use the "middle point" $f(x_0)$, since $c_0 = 0$).

3. Some Useful Notation

Throughout this course, we use the following notation: for $a, b \in \mathbb{R}$ with $a < b$,

(1) (a, b) is the (open) interval $\{x \in \mathbb{R} \mid a < x < b\}$;
(2) $[a, b]$ is the (closed) interval $\{x \in \mathbb{R} \mid a \le x \le b\}$;
(3) $C[a, b]$ is the set of $f\colon [a, b] \to \mathbb{R}$ that are continuous on $[a, b]$;
(4) $C(a, b)$ is defined similarly;
(5) for $k \in \mathbb{N} = \{1, 2, \ldots\}$, $C^k[a, b]$ denotes the set of functions $[a, b] \to \mathbb{R}$ that are k times continuously differentiable on $[a, b]$ (for the endpoints a, b we take the derivative only from one side);
(6) $C^k(a, b)$ is defined similarly;
(7) $C^\omega(a, b)$ is the set of *(real) analytic* functions $f\colon (a, b) \to \mathbb{R}$, i.e., such that for each $x_0 \in (a, b)$, for $|h|$ sufficiently small, the Taylor series for $f(x_0 + h)$ converges and equals $f(x_0 + h)$, i.e.,

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

and the right-hand-side converges for $|h|$ sufficiently small.

4. ODE's (Ordinary Differential Equations)

ODE's (Ordinary Differential Equations) are studied in Chapter 16 of [A&G].

4.1. **Notation in [A&G]: Univariate ODE's.** For a single variable ODE, they use the notation (Section 16.1, page 481)

$$y' = \frac{dy}{dt} = f(t, y), \quad a \le t \le b,$$

and one wishes to determine all solutions $f\colon (a, b) \to \mathbb{R}$ to this differential equation. Since $y = y(t)$ depends on t (where in dynamics t often denotes "time"), the above equation really reads

$$y'(t) = f\big(t, y(t)\big).$$

The *initial value problem* imposes the condition

$$y(a) = c$$

for some $a, c \in \mathbb{R}$, and one seeks a (hopefully unique) solution $f\colon (a, b) \to \mathbb{R}$ for some real $b > a$. Outside of numerical analysis it is far more common to write

$$y(t_0) = y_0,$$

where t_0 (or a above) represents the *initial time*, and y_0 (or c above) is the *initial value of* $y = y(t)$. The reason that [A&G] uses a instead of t_0 is the common theme in their textbook, including the chapters on polynomial interpolation, splines, and integration, where the interval of interest is $[a, b] \subset \mathbb{R}$.

One also commonly uses the abbreviation $\dot{y}$ for y' or dy/dt, especially in celestial mechanics.

We also remark that the condition $y(a) = c$ or $y(t_0) = y_0$ often determines a unique solution $y = y(t)$ defined for all $t < t_0$ with t sufficiently close to t (and sometimes all real $t < t_0$).

4.2. **Notation in [A&G]: Multivariate (or m-dimensional) ODE's.** For a "system of m ODE's," with $m \in \mathbb{N}$, [A&G] use the notation

$$\mathbf{y}' = \frac{d\mathbf{y}}{dt} = \mathbf{f}(t, \mathbf{y}), \quad a \leq t \leq b,$$

where $\mathbf{y} \colon (a, b) \to \mathbb{R}^m$; the above is really shorthand for the equation

$$\mathbf{y}'(t) = \mathbf{f}\big(t, \mathbf{y}(t)\big), \quad a \leq t \leq b;$$

the *initial value problem* refers to finding a $y = y(t)$ defined for $a \leq t \leq b$, or at least $a \leq t \leq a + \epsilon$ for some $\epsilon > 0$, subject to

$$\mathbf{y}(a) = \mathbf{c}$$

for a given $\mathbf{c} \in \mathbb{R}^m$. Notice that for the "one-body problem," m typically denotes "mass," and hence we will often write n (or some other letter) in place of m.

It turns out that many of the same principles for a single variable ODE hold for general ODE's.

In class we typically write the initial value problem as

$$\mathbf{y}' = \mathbf{f}(t, \mathbf{y}), \quad \mathbf{y}(t_0) = \mathbf{y}_0$$

for given $t_0 \in \mathbb{R}$ and $\mathbf{y}_0 \in \mathbb{R}^n$, and when discussing physics, specifically celestial mechanics, we write

$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}), \quad \mathbf{y}(t_0) = \mathbf{y}_0.$$

4.3. **Common Mathematics Notation and Sources for ODE's.** I highly recommend UBC Math's second-term calculus book: https://personal.math.ubc.ca/~CLP/CLP2/clp_2_ic_text.pdf, by Feldman, Rechnitzer, and Yeager, including Section 2.4 as in introduction to ODE's, https://personal.math.ubc.ca/~CLP/CLP2/clp_2_ic_text.pdf#page=249&zoom=100,94,861, and Appendix D on Numerical Solutions to ODE's: https://personal.math.ubc.ca/~CLP/CLP2/clp_2_ic_text.pdf#page=435&zoom=100,94,78.

The above textbook, as well as many other mathematics textbooks, use the notation of [A&G], where x is used in place of t, i.e.,

$$y'(x) = \frac{dy}{dx}(x) = f\big(x, y(x)\big), \quad y(x_0) = y_0$$

for a single-variable ODE, where $y(x_0) = y_0$ and you wish to solve this for a function $y \colon \mathbb{R} \to \mathbb{R}$, or at least $y \colon (a, b) \to \mathbb{R}$ where $a < x_0 < b$, and a is as small as possible, and b is as large as possible. Of course, in the "initial value problem" we are often—in practice—mainly interested in $y = y(x)$ for $x \geq x_0$.

As soon as you learn integration, you can solve many *separable ODE's*, meaning ODE's of the form

$$y'(x) = h(x)g\big(y(x)\big)$$

(the UBC textbook [FRY] writes $y' = f(x)g(y)$ with f replacing h above). The idea is to write this as

$$\frac{dy}{g(y)} = h(x)dx$$

and to "integrate"; this doesn't always work, but it does if $g(y)$ is differentiable near $y = y_0$ and $h(x)$ is continuous near $x = x_0$. We will give a number of examples below.

4.4. **Common Celestial Mechanics Notation.** Celestial mechanics often uses the notation $\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x})$. Historically Newton explained Kepler's observations by the differential equation

$$m\ddot{\mathbf{x}} = -GMm\frac{\mathbf{x}}{\|\mathbf{x}\|_2^3} \tag{1}$$

(expressing "mass times acceleration equals force"), where

(1) M is the mass of the Sun, viewed as located at the origin $\mathbf{0} = (0,0) \in \mathbb{R}^2$; and
(2) m is the mass of a planet, whose position $\mathbf{x} = \mathbf{x}(t) = (x_1(t), x_2(t))$ is a function $\mathbb{R} \to \mathbb{R}^2$; and
(3) G is a universal *Gravitational constant*.

Hence the Sun pulls the planet toward the Sun—in the direction of $-\mathbf{x}$—by a force proportional to $1/\|\mathbf{x}\|_2^2$.

For reasons that will become clear below, given the "initial conditions"

$$\mathbf{x}(t_0) = \mathbf{x}_0, \quad \mathbf{v}(t_0) = \mathbf{v}_0,$$

one easily sees that $\mathbf{x}(t)$ lies entirely in the two-dimensional place spanned by $\mathbf{x}_0, \mathbf{v}_0$ (even if $\mathbf{x} = \mathbf{x}(t)$ takes values in $\mathbb{R}^d$ for $d \geq 3$). It will follow that one may as well assume that $\mathbf{x}(t)$ takes values in $\mathbb{R}^2$.

One reduces this to the form $\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y})$ by setting $\mathbf{v} = \dot{\mathbf{x}}$ (the velocity), $\mathbf{a} = \dot{\mathbf{v}} = \ddot{\mathbf{x}}$ (the acceleration), and we write "mass times acceleration equals Force" as

$$m\ddot{\mathbf{x}} = m\mathbf{a} = \mathbf{F} = -GMm\mathbf{x}/\|\mathbf{x}\|_2^3.$$

Writing $g = GM$ and $\mathbf{y} = (\mathbf{v}, \mathbf{x}) = (v_1, v_2, x_1, x_2)$, we get a "system of 4 ODE's":

$$\frac{d}{dt}\begin{bmatrix} v_1 \\ v_2 \\ x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -gx_1/r^3 \\ -gx_2/r^3 \\ v_1 \\ v_2 \end{bmatrix}, \quad \text{where } r = \|\mathbf{x}\|_2 = \sqrt{x_1^2 + x_2^2}.$$

In other words, we set

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \\ x_1 \\ x_2 \end{bmatrix}$$

and we have

$$\dot{\mathbf{y}} = \begin{bmatrix} -gy_3/(y_3^2+y_4^2)^{3/2} \\ -gy_4/(y_3^2+y_4^2)^{3/2} \\ y_1 \\ y_2 \end{bmatrix}.$$

4.5. **Force Equals Mass Times Acceleration.** A vast class of ODE's arise from the formula "force equals mass times acceleration."

The simplest form of these equations assumes there are n rigid bodies, whose masses are $m_1, \ldots, m_n$, and whose centre of masses are functions $\mathbf{x}_1 = \mathbf{x}_1(t), \ldots, \mathbf{x}_n(t)$ which are moving in $\mathbb{R}^d$, so that each $\mathbf{x}_i = \mathbf{x}_i(t)$ is a function $\mathbb{R} \to \mathbb{R}$, or at times $[a,b] \to \mathbb{R}$, where $a < b$ and $t = a$ is the "initial time," and $t = b$ is the "final time" that interests us.

We often of the masses as "point masses," meaning that the entire i-th mass is "concentrated at the $\mathbf{x}_i = \mathbf{x}_i(t)$;" this idealized situation often simplifies the laws of physics involved.

For point masses, the formula "force equals mass times acceleration" implies that

$$m_i \ddot{\mathbf{x}}_i = m_i \mathbf{x}_i''(t) = \mathbf{F}_i(t),$$

where $\mathbf{F}_i(t)$ is the force exerted upon the i-th body at time t. Typically the function $\mathbf{F}_i(t)$ is far too complicated to write down; however, one can often write

$$\mathbf{F}_i(t) = \mathbf{f}_i\big(t, \mathbf{x}_1(t), \mathbf{x}_1'(t), \ldots, \mathbf{x}_n(t), \mathbf{x}_n'(t)\big),$$

where f_i can be written more simply, or even just

$$\mathbf{F}_i(t) = \mathbf{f}_i\big(\mathbf{x}_1(t), \ldots, \mathbf{x}_n(t)\big).$$

For example, if you toss an apple on the Earth's surface, and its centre of mass is therefore a function $f\colon [t_0, t_{\text{end}}] \to \mathbb{R}^3$, its equation of motion is often reasonably modelled as

$$m\mathbf{x}''(t) = \mathbf{f}\big(\mathbf{x}, \mathbf{x}'\big) = (0, 0, -gm) - \mu(\|\mathbf{x}'\|)\mathbf{x}',$$

where $(0,0,-gm)$ is an approximation of the Earth's gravity on its surface (so g is roughly $9.807\,\text{metres}^2/\text{sec}$), and $\mu\colon [0,\infty) \to [0,\infty)$ is a non-negative function defined on the non-negative reals which represents a wind resistance force. Neglecting wind resistance we get

$$m\mathbf{x}''(t) = (0, 0, -gm),$$

which is often a reasonable approximation, as a function $[t_0, t_{\text{end}}]$, where t_0 is the time at which the apple is tossed, and t_{end} might be when the apple hits something and breaks into smaller pieces.

Remark 4.1. In class in 2024, I slid a coffee travel mug along a table to illustrate that it can come to a halt (primarily due to the friction with the table). However, you wouldn't expect to see the same thing happen if time were reversed (i.e., the mug at rest begin to move toward my hand). Hence ODE's from physics that involve forces arising from $\mathbf{x}'$ typically don't follow the same laws of physics if time were reversed. By contrast, one can typically "reverse time" of any solution to $\mathbf{x}'' = \mathbf{f}(\mathbf{x})$ to get another valid solution. This was indicated in Homework of 2024.

4.6. **The One-Body and Central Force Problems.** The first few classes begin with an example of the "3-body problem," meaning, three celestial bodies (stars, planets, etc.) of the same mass, all moving in a plane.

It is much simpler to study the 2-body problem, which can often be reduced to a "1-body force field" problem. In this case (1) is generalized to

$$m\ddot{\mathbf{x}} = -\mu(\|\mathbf{x}\|_2)\frac{\mathbf{x}}{\|\mathbf{x}\|_2}, \tag{2}$$

which is called a *central force problem*, where $\mu\colon [0,\infty] \to \mathbb{R}$ often takes only positive values to represent a single point mass that has a force of magnitude $\mu(\|\mathbf{x}\|_2)$ pulling it toward the origin $\mathbf{0} \in \mathbb{R}^d$ (negative values of μ indicate a repulsive force from $\mathbf{0}$). One similarly writes $\mathbf{v} = \dot{\mathbf{x}}$ and forms a $2d$-dimensional ODE (or "a system of $2d$ ODE's"). It follows that the initial value problem in this case specifies $t_0 \in \mathbb{R}$ and $\mathbf{x}_0, \mathbf{v}_0 \in \mathbb{R}^d$, and seeks a unique solution to (1) with

$$\mathbf{x}(t_0) = \mathbf{x}_0, \quad \mathbf{v}(t_0) = \mathbf{v}_0$$

(as remarked earlier, one easily sees that $\mathbf{x}(t)$ always lies in the plane (or line or $\mathbf{0}$) spanned by $\mathbf{x}_0, \mathbf{v}_0$, hence one can reduce this to the case $d = 2$). It is easy to see that $\mathbf{v}(t_0) = \mathbf{v}_0 = \mathbf{0}$, then in some finite time t_1, the limit of $\mathbf{x}(t)$ as $t \to t_1$ is $\mathbf{0}$, and the speed $\|\mathbf{v}(t)\|$ is faster than the speed of light for $t < t_1$ with t sufficiently close to t_1. Moreover at $t = t_1$ the equation becomes "singular," and the equation—at least as written—make no sense at $t = t_1$. [This turns out to be a "mild singularity" as far as singularities go.]

The more general "one-body" or "force field" equation does not assume that the force acts in the direction from $\mathbf{x}$ to $\mathbf{0}$ and does not assume the force's magnitude depends only on $\|\mathbf{x}\|_2$. Hence this problem can be written as

$$m\ddot{\mathbf{x}} = \mathbf{f}(\mathbf{x}).$$

An especially interesting case of this equation is the situation where there is a "potential," meaning a function $\phi\colon \mathbb{R}^n \to \mathbb{R}$ such that $\nabla\phi = \mathbf{f}$ (we will explain what this means when we discuss partial derivatives). In this case, using vector calculus it is easy to see that the "energy"

$$E(t) \stackrel{\text{def}}{=} (1/2)\|\mathbf{v}(t)\|^2 - \phi\big(\mathbf{x}(t)\big)$$

is independent of time (in brief, the equation $\|\mathbf{v}(t)\|^2 = \mathbf{v}\cdot\mathbf{v}$, and the chain rule and the product rule imply that $\dot{E} = m\mathbf{v}\cdot\mathbf{a} - (\nabla\phi)(\mathbf{x})\cdot\mathbf{v} = 0$, by the equation $m\mathbf{a} = \mathbf{f}(\mathbf{x}) = \nabla\phi(\mathbf{x})$). Once commonly refers to $(1/2)\|\mathbf{v}(t)\|^2$ as the *kenetic energy* and to $\phi\big(\mathbf{x}(t)\big)$ (or, at times, $-\phi$) as the *potential energy*. (Note that ϕ is only well-defined up to an additive constant, even thought there is sometime a natural choice of ϕ. Hence energy is only defined up to an additive constant, as is typically the case in physics.)

Example 4.2. The central force problem is a special case of the one-body problem where the force field $\mathbf{f}(\mathbf{x})$ has a potential, namely, $\phi = \nu(\mathbf{x})$, where

$$\nu(r) = \int -\mu(r)\,dr$$

(which is an indefinite integral). So

(1) for Newton's law $\mu(r) = 1/r^2$, $\nu(r) = 1/r$ plus an arbitrary constant, although one usually takes $\nu(r) = 1/r$ (so that at "infinite distance" the potential energy is 0);
(2) for the harmonic oscillator (see the next section), $\mu(r) = Cr$, we have $\nu(r) = Cr^2/2$.

Example 4.3. The force field on the Earth's surface of roughly $(0, 0, -gm)$, with g is roughly 9.807 metres^2/sec), has a potential $f(\mathbf{x}) = x_3 g + C$ for an arbitrary constant C.

Example 4.4. Certain swimming complexes have a "kid's ring pool" that is an annulus, with, say, a gentle[2] and counterclockwise force (looking from above, which is common) due to an externally produced the motion of the water (and the water's viscosity). In this case, it is more difficult to move clockwise than counterclockwise, and hence to move 180° around the pool, it takes more work to move clockwise than counterclockwise. This is an example of a force field where the work needed from point A to point B depends on the path you take; such force fields are never potential fields. Moreover, such a force field has a non-zero "curl:" so if the force field restricted to the annulus is $(x_2, -x_1, 0)$ (for simplicity), then $\nabla \times (x_2, -x_1, 0) = (0, 0, -2)$ (looking from above) using Hamilton's quaternionic definition of $\times$ (or $(0, 0, 2)$ looking from below, which reverses the z-axis direction, and thereby the sign of "$\hat{k}$" following Hamilon's convention based on $ij = k$, rather than the equally possible convention $ij = -k$). By constrast, any potential force field $\mathbf{F}(\mathbf{x}) = \nabla(\phi(\mathbf{x}))$ has "curl" equal to $\nabla \wedge (\nabla(\phi(\mathbf{x}))$, which we easily see must vanish.

Remark 4.5. As suggested in the above example, the amount of work it takes to move from point A to point B in a potential force field $\mathbf{F}(\mathbf{x}) = (\nabla\phi)(\mathbf{x})$ is independent of the path you take; this results from the fact that if $\mathbf{p}\colon [a, b] \to \mathbb{R}^d$ is differentiable, and $\mathbf{p}(a) = A$ and $\mathbf{p}(b) = B$, then, using the chain rule (for the first equality below),

$$\int_{s=a}^{s=b} (\nabla\phi)(\mathbf{p}(s)) \cdot \frac{d\mathbf{p}}{ds}(s)\, ds = \int_{s=a}^{s=b} \frac{d}{ds}\big(\phi(\mathbf{p}(s)\big)\, ds = \phi(\mathbf{p}(b)) - \phi(\mathbf{p}(a)) = \phi(B) - \phi(A).$$

4.7. **Introduction to the n-Body Problem.** The first few classes begin with an example of the "3-body problem," meaning, three celestial bodies (stars, planets, etc.) of the same mass, all moving in a plane.

Here we want to introduce some basic aspects of the n-body problem. We imagine n bodies that lie in d-dimensional space, so for each $i = 1, \ldots, n$, the i-body has mass m_i and position $\mathbf{x}_i = \mathbf{x}_i(t)$ where $\mathbf{x}_i\colon \mathbb{R} \to \mathbb{R}^d$ assuming that we have a solution valid for all time t.

(Otherwise maybe $\mathbf{x}_i\colon [t_0, t_{\text{end}}] \to \mathbb{R}^d$, where t_0 is an initial time, and t_{end} is and ending time that interests us; [A&G] often uses $[a, b]$ instead of $[t_0, t_{\text{end}}]$, for consistency with previous topics, such as interpolation or quadrature.)

It is common to write

$$\mathbf{v}_i(t) = \dot{\mathbf{x}}(t), \quad \mathbf{a}_i(t) = \dot{\mathbf{a}}(t) = \ddot{\mathbf{x}}(t)$$

for the velocity and acceleration of the i-th body.

[2]Gentle for typical pools and typical volicities of kids and their (grand)parents.

First we assume that "force equals mass times acceleration" yields the equations of motion

$$m_i \mathbf{x}_i(t) = \sum_{j \neq i} \mathbf{F}_{ij}(t),$$

where $\mathbf{F}_{ij}(t)$ is the force that the j-th body exerts on the i-th body.

Proposition 4.6. *Let $n, d \in \mathbb{N}$ with $n \geq 1$. Say consider any solution of the system $m_i \mathbf{x}_i(t) = \sum_{i \neq j} \mathbf{F}_{ij}(t)$ for $i = 1, \ldots, n$ with $\mathbf{x}_i(t)$ taking values in $\mathbb{R}^d$. such that for all $i \neq j$ we have $\mathbf{F}_{ij}(t) = -\mathbf{F}_{ji}(t)$ for all t. Then the* total momentum

$$M(t) \stackrel{def}{=} \sum_{i=1}^n m_i \mathbf{v}_i(t) = \sum_{i=1}^n m_i \dot{\mathbf{x}}_i(t)$$

is constant. If, moreover, $\mathbf{F}_{ij}(t) = -\mathbf{F}_{ji}(t)$ is colinear with $\mathbf{x}_i(t) - \mathbf{x}_j(t)$, then the angular momentum*,*

$$\mathbf{\Omega}(t) \stackrel{def}{=} \sum_{i=1}^n m_i \mathbf{x}_i(t) \wedge \mathbf{v}_i(t)$$

is constant; if you don't know what $\wedge$ means and $d \leq 3$, there is no harm in viewing $\mathbb{R}^d$ as equal to or a subspace of $\mathbb{R}^3$ as usual, and defining $(x_1, x_2, x_3) \times (v_1, v_2, v_3)$ to be the "vector part" of $(x_1 i + x_2 j + x_3 k)(v_1 i + v_2 j + v_3 k)$ under Hamilton's quaternionic convention $ij = k$ (and $i^2 = j^2 = k^2 = -1$).

Proof. We have

$$\dot{M}(t) = \sum_{i=1}^n m_i \ddot{\mathbf{x}}(t) = \sum_{i \neq j} \mathbf{F}_{ij}(t) = \sum_{i<j} \big(\mathbf{F}_{ij}(t) + \mathbf{F}_{ji}(t)\big) = \sum_{i<j} \mathbf{0},$$

and hence $\dot{M}(t) = \mathbf{0}$ and hence $M(t)$ is constant. Using the fact that $\wedge$ (or $\times$) here is a bilinear, anti-symmetric operator, the product rule also applies to $\wedge$, and hence (suppressing the t's):

$$\dot{\mathbf{\Omega}} = \sum_{i=1}^n m_i \Big(\dot{\mathbf{x}}_i \wedge \mathbf{x}_i + \dot{\mathbf{x}}_i \wedge \ddot{\mathbf{x}}_i \Big) = \sum_{i=1}^n m_i \mathbf{x}_i \wedge \left(\sum_{i<j} \mathbf{F}_{ij} \right) = \sum_{i<j} (\mathbf{x}_i - \mathbf{x}_j) \wedge \mathbf{F}_{ij} = \mathbf{0},$$

and hence $\mathbf{\Omega} = \mathbf{\Omega}(t)$ is contant. □

Remark 4.7. We remark that for $d \geq 4$, $\wedge$ takes two vectors in $\mathbb{R}^d$ and returns a value in a space of dimension $\binom{d}{2} = d(d-1)/2$ (that is bilinear and anti-symmetric in its two arguments) Hamilton's deserved fascination with quaternions influenced many others, including Maxwell[3] lead to Hamilton's definition of $\times$ as the "vector part" of quaternionic multiplication, which was profoundly important in the history of physics. However, confusing $\wedge$ with $\times$ can, at times, can cause needless confusion.

When we discuss partial derivatives, we will be able to prove the usual "conservation of Energy" when the forces involved are potential forces.

[3] Maxwell's famous equations first appeared in [**?**], pages blah. On page blah Maxwell write that they doens't want to assume familiarity with quaternions, but on pages blah they explains how to simplify notation using Hamilton's quaternions. We thank Prof. Daniel Shapiro for discussions on regarding Hamilton and Maxwell.

Proposition 4.8. *Say that in Proposition 4.6, for each $1 \le i < j \le m$ there is a function $\phi_{ij}\colon \mathbb{R}^d \to \mathbb{R}$ such that*

$$F_{ij}(t) = \nabla \phi_{ij}(\mathbf{x}_j - \mathbf{x}_i),$$

Then (the total energy *of the system, namely)*

$$E \stackrel{def}{=} \left(\sum_{i=1}^{n} (1/2) m_i \|\dot{\mathbf{x}}_i\|_2^2\right) - \left(\sum_{i<j} \phi_{ij}\big(\mathbf{x}_j - \mathbf{x}_i\big)\right)$$

is constant in time.

CHECK THE SIGNS IN THE ABOVE...

Remark 4.9. When using an ODE solver to generate approximate solutions, it can be useful to have time-invariant quantities such as momentum, angular momentum, and total energy. If these quantities are not conserved numerically, then this can indicate trouble. However, the converse isn't true unless all conserved quantities and other inferrable quantities determine the $\mathbf{x}_i(t)$. This is true for the two-body problem under Newtonian gravitation (or the analogous one-body problem), but this isn't generally true; even for the this two-body problem, the invariants involve fairly complicated functions (but functions that are explicit, in terms of polynomials and sines/cosines).

5. Basic Theory and Examples of ODE's

The basic existence and uniqueness theory for ODE's can be understood from the one-variable ODE $y' = f(t, y)$ (we now use the notation in [A&G], where $y' = dy/dt$ and $y = y(t)$ is a function of t). In fact, the basic existence and uniqueness (and a number of examples) can be understood when $f(t, y) = f(y)$ is independent of t; this is often the case, with physical laws that don't change in time. Hence we will start there.

5.1. The ODE $y' = f(y)$.

This ODE $y' = f(y)$ is therefore separable, and one can often solve it by writing

$$\frac{dy}{dt} = f(y) \quad \Rightarrow \quad \int \frac{dy}{f(y)} = \int dt = t + C,$$

and evaluating the indefinite integral above. This always works when f is a differentiable function of y; this can be seen from the basic existence and uniqueness theorem below.

5.2. The Basic Existence and Uniqueness Theorem for $y' = f(y)$.

Theorem 5.1. *Let $t_0, y_0 \in \mathbb{R}$, and let $f = f(y)$ be a function that is defined in a neighbourhood of y_0 (i.e., $f\colon (y_0 - \delta, y_0 + \delta) \to \mathbb{R}$ for some $\delta > 0$). Then the ODE*

$$y' = f(y), \quad y(t_0) = y_0, \tag{3}$$

1. *has a local solution, i.e., a solution $y = y(t)$ defined for $t \in (t - \epsilon, t + \epsilon)$ for $\epsilon > 0$ if $f \in C^0(y_0 - \delta, y_0 + \delta)$, i.e., $f = f(y)$ is continuous in $(y_0 - \delta, y_0 + \delta)$;*
2. *moreover the solution is unique if $f(y)$ is Lipschitz continuous in $(y_0 - \delta, y_0 + \delta)$, i.e., $|f(y_2) - f(y_1)| \le K|y_2 - y_1|$ for some K whenever $y_1, y_2 \in (y_0 - \delta, y_0 + \delta)$ (this is true whenever f' exists in $(y_0 - \delta, y_0 + \delta)$ and $|f'(y)| \le K$ in this interval;*

(3) *if* $f \in C^k(y_0-\delta, y_0+\delta)$ *for some* $k = 0, 1, \dots$ *or* $k = \infty$ *or* $k = \omega$, *then* $y \in C^{k+1}(t_0-\epsilon', t_0+\epsilon')$ *for some* $\epsilon' > 0$ *(where* $\infty+1 = \infty$ *and* $\omega+1 = \omega$*);*

(4) *if* $f \in C^0(\mathbb{R})$ *(i.e.,* $f\colon \mathbb{R} \to \mathbb{R}$ *is everywhere continuous) and for all* $y \in \mathbb{R}$ *we have* $|f(y)| \le K_1|y| + K_2$ *for some constants* $K_1, K_2 \ge 0$, *then any solution to* (3) *is defined for all* $t \ge t_0$, *and there we have*

$$|y(t)| \le z(t),$$

where $z(t)$ *is the solution to* $z' = K_1 z + K_2$ *and* $z(t_0) = |y_0|$, *i.e.,* $z(t) = (Ce^{K_1} - K_2)/K_1$ *where* C *satisfies* $Ce^{K_1} - K_2 = K_1|y_0|$, *i.e.,* $C = (K_1|y_0| + K_2)e^{-K_1}$.

The proof of each part of this theorem represents some fundamental ideas regarding ODE's; we will outline the proofs in an appendix.

5.3. **The Integral form of the ODE.** It is important to understand the *integral form* of (3); this is obtained by writing

$$y(t) - y(t_0) = \int_{s=t_0}^{s=t} y'(s)\, ds$$

and hence $y' = f(y)$ implies that

$$y(t) = y(t_0) + \int_{s=t_0}^{s=t} f\big(y(s)\big)\, ds; \tag{4}$$

this "integral form" will be used to prove Theorem 5.1 and to derive numerical (approximate) solutions to the ODE in (3)).

Intuitively speaking, the ODE (3) assumes that y is differentiable, whereas (4) only assumes that one can make sense of the integral of $f(y(s))$; the integral form can be used to make sense of this ODE when f is not even continuous (or could be a generalized function, e.g., involving the "Dirac" delta function). Moreover, if we have a function $y_{\text{approx}}(t)$ that is an approximation of a true solution, then the function

$$\Phi(y_{\text{approx}})(t) \stackrel{\text{def}}{=} y(t) = y(t_0) + \int_{s=t_0}^{s=t} f\big(y_{\text{approx}}(s)\big)\, ds$$

turns out to be a "better approximation" of a true solution (in a number of senses, for t near t_0; see Appendix **??**). One consequence is that if we start with any function $y_0(t)$, and let $y_1 = \Phi(y_0)$, $y_2 = \Phi(y_1)$, etc., then the "iterates of Φ on y_0," i.e., the sequence $y_0, y_1, y_2, \dots$ will converge to a local solution of the ODE (3) when f is Lipschitz continuous.

5.4. **Basic Examples of $y' = f(y)$.**

Example 5.2. The solution $y' = Ay$ was "guessed" in class to be $y(t) = e^{A(t-t_0)}y_0$ (this can be done simply by "guessing," or by writing $dy/dt = Ay$, hence $dy/y = A dt$ and integrating, which is really a more systematic way of guessing). Since this equation is of the form $y' = f(y)$ where $f(y) = Ay$, Theorem 5.1 can be used to show that this is the unique solution to (3). Note that this solution $y = y(t)$ exists for all $t \in \mathbb{R}$, not merely all $t \ge t_0$.

Example 5.3. The solution to $y' = y^2$ and $y(1) = 1$ was shown in class to be $y(t) = 1/(2-t)$, and hence $y(t) \to \infty$ as $t \to 2-$ (i.e., $t < 2$ and $t \to \infty$). Hence solutions to ODE's can tend to infinity at some finite time, when $y' = f(y)$ and $f(y)$ grows faster than linear in y; by contrast, this doesn't happen if $|f(y)| \le K_1|y| + K_2$.

Example 5.4. In class we showed that the solution to $y' = |y|^{1/2}$ has the following solution for any $a, b \in \mathbb{R}$ with $a < b$:

$$y(t) = \begin{cases} -(1/4)(t-a)^2 & \text{if } t \le a, \\ 0 & \text{if } a \le t \le b, \\ (1/4)(t-b)^2 & \text{if } t \ge b. \end{cases}$$

It is not hard to show that such $y(t)$ are the only possible solutions (if we allow for the "limiting cases" $a = -\infty$ and $b = +\infty$), since $f(y) = |y|^{1/2}$ is infinitely differentiable for $y \neq 0$. We easily see that $y'(t)$ exists for all t and that

$$y'(t) = \begin{cases} -(1/2)(t-a) & \text{if } t \le a, \\ 0 & \text{if } a \le t \le b, \\ (1/2)(t-b) & \text{if } t \ge b. \end{cases}$$

It follows that $y''(a-)$, i.e., "y'' at a from the left" (or "left derivative of y' at a) equals

$$y''(a-) = \lim_{t \to a-} \frac{y'(a) - y'(t)}{a - t} = \lim_{t \to a-} \frac{0 - (-1/2)(t-a)}{a-t} = \lim_{t \to a-} \frac{-1}{2} = -1/2,$$

and similarly

$$y'(b+) = 1/2,$$

and if $a < b$ then $y'(a+) = y'(b-) = 0$. Hence y'' does not exist at a and b (assuming both a, b are finite, regardless of whether or not $a = b$ or $a < b$). Hence $y(t)$ is always differentiable, but *never* twice differentiable (unless we consider the limiting case $a = -\infty$ and $b = +\infty$, meaning $y(t) = 0$ for all t). In particular, the initial value problem

$$y' = |y|^{1/2}, \quad y(t_0) = y_0$$

never has a unique solution. We will also see that this ODE has some rather "chaotic" behaviour when we try to solve it numerically.

5.5. **m-dimensional ODE's.** A lot of the theory, and some simple examples of ODE's (e.g., those with constant coefficients) generalize almost word-for-word to m-dimensional ODE's and the initial value problem

$$\mathbf{y}'(t) = \mathbf{f}\big(t, \mathbf{y}(t)\big), \quad \mathbf{y}(t_0) = \mathbf{y}_0.$$

For example the constant coefficient ODE $\mathbf{y}' = A\mathbf{y}$, hence $\mathbf{y}(t)$ has values in $\mathbb{R}^m$ and $A \in \mathbb{R}^{n \times n}$, i.e., A is a real $n \times n$ matrix. Its general solution is

$$\mathbf{y}(t) = e^{A(t-t_0)}\mathbf{y}_0.$$

In the next section we elaborate on this and explain what we mean by $e^{A(t-t_0)}$ when A is a square matrix. (The same is true if $A \in \mathbb{C}^{m \times m}$ and $\mathbf{y}(t)$ has values in $\mathbb{C}^m$; here t is (most simply) still real-valued.)

Similarly Theorem 5.1 generalizes to the m-dimensional case, with etc.

ADD MORE HERE?

6. The Harmonic Oscillator and the ODE $\mathbf{y}' = A\mathbf{y}$

Consider a function $x \colon \mathbb{R} \to \mathbb{R}$ (or, at times, $x \colon [a, b] \to \mathbb{R}$ for some given reals $a < b$), which represents the centre of mass of some object moving along the x-axis; it is often simplest to think of the object as a "point mass," meaning a mass entirely concentrated at a single point.

So say a point mass moves entirely along the x-axis, whose position is $x = x(t)$, and is subject to a force exerted on it by a spring (with one end connected to the point mass, and another end anchored somewhere along the x-axis). In class we explained that if the "rest position" of the end of the spring connected to the point mass is x_0, the equation "force equals mass times acceleration" is often modelled as

$$mx'' = -C(x - x_0)$$

where $C > 0$ is a constant depending on the spring. By translating the x coordinate we can assume that $x_0 = 0$, i.e., $mx'' = -Cx$. Scaling time we may restrict to $x'' = -x$. Hence, writing $v = x'$ (which is the velocity of the point mass), we have the 2-dimensional system

$$\frac{d}{dt}\begin{bmatrix} v \\ x \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} v \\ x \end{bmatrix},$$

or

$$\mathbf{y}' = A\mathbf{y}, \quad \text{where} \quad \mathbf{y} = \begin{bmatrix} v \\ x \end{bmatrix}, \quad A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

Given the solution to the one-dimensional ODE $y' = Ay$, we could guess that the solution to the above ODE, subject to the initial value

$$\mathbf{y}(t_0) = \mathbf{y}_0, \quad \text{i.e.,} \quad \begin{bmatrix} v(t_0) \\ x(t_0) \end{bmatrix} = \begin{bmatrix} v_0 \\ x_0 \end{bmatrix},$$

for $v_0, x_0, t_0 \in \mathbb{R}$ might be

$$\mathbf{y}(t) = e^{A(t-t_0)}\mathbf{y}_0,$$

where for a square matrix, M, we define

$$e^M = I + M + (1/2)M^2 + (1/3!)M^3 + \cdots$$

Of course, we have to convince ourselves that this infinite series makes sense, and that we can differentiate the series

$$e^{A(t-t_0)} = I + A(t-t_0) + (1/2)\big(A(t-t_0)\big)^2 + (1/3!)\big(A(t-t_0)\big)^3 + \cdots$$

term by term, so that

$$\begin{aligned}\Big(e^{A(t-t_0)}\Big)' &= \Big(I + A(t-t_0) + (1/2)\big(A(t-t_0)\big)^2 + (1/3!)\big(A(t-t_0)\big)^3 + \cdots\Big)' \\ &= I' + A(t-t_0)' + (1/2)A^2\big((t-t_0)^2\big)' + (1/3!)A^3\big((t-t_0)^3\big)' + \cdots \\ &= A + (1/2)A^2 2(t-t_0) + (1/3!)A^3 3(t-t_0)^2 + \cdots = Ae^{A(t-t_0)},\end{aligned}$$

which mimics the computation when A is a 1×1 matrix, i.e., a real number.

To compute $e^{A(t-t_0)}$ for the above case of A, we notice that

$$A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \; A^2 = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}^2 = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} = -I,$$

and therefore $A^4 = (-I)^2 = I$, and hence

$$I = A^0 = A^4 = A^8 = \cdots, \; A = A^5 = A^9 = \cdots,$$

etc., and hence

$$e^{A(t-t_0)} = \begin{bmatrix} 1 - (1/2)(t-t_0)^2 + (1/4!)(t-t_0)^4 - \cdots & -(t-t_0) + (1/3!)(t-t_0)^3 - (1/5!)(t-t_0)^5 + \cdots \\ (t-t_0) - (1/3!)(t-t_0)^3 + (1/5!)(t-t_0)^5 - \cdots & 1 - (1/2)(t-t_0)^2 + (1/4!)(t-t_0)^4 - \cdots \end{bmatrix}$$

$$= \begin{bmatrix} \cos(t-t_0) & -\sin(t-t_0) \\ \sin(t-t_0) & \cos(t-t_0) \end{bmatrix}$$

7. Euler's Method and the (Explicit) Trapezoidal Rule

Sections 16.2 of [A&G] discusses Euler's method to approximate solutions of ODE's, and 16.3 gives a much wider class of ODE approximation methods known as Runge-Kutta methods, which includes Euler's method and a common improvement of Euler's method known as the (explicit) trapezoidal rule.

7.1. **Euler's Method.** Euler's method to numerically approximate a solution to the initial value problem

$$\mathbf{y}' = \mathbf{f}(t, \mathbf{y}), \quad \mathbf{y}(t_0) = \mathbf{y}_0,$$

can be derived as follows: for "small $h > 0$" we have

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

and therefore

$$\mathbf{y}(t+h) \approx \mathbf{y}(t) + h\mathbf{y}'(t) = \mathbf{y}(t) + h\mathbf{f}\big(t, \mathbf{y}(t)\big). \tag{5}$$

Hence, choose a small $h > 0$ (which depends on various considerations, including how much compute power you have); the setting

$$t_1 = t_0 + h, \ t_2 = t_0 + 2h, \ldots,$$

i.e., each $i \in \mathbb{Z}$, $t_i = t_0 + ih$, we can approximate $y(t_i)$ by applying (5) with $t = t_i$ to obtain

$$\mathbf{y}(t_{i+1}) = \mathbf{y}(t_i + h) \approx y(t_i) + hf\big(t_i, y(t_i)\big).$$

This produces $y_1, y_2, \ldots$, which respectively approximate $y(t_1), y(t_2), \ldots$, given recursively as

$$y_{i+1} = y_i + hf(t_i, y_i).$$

We may similarly produce $y_{-1}, y_{-2}, \ldots$ approximating $y(t_{-1}), y(t_{-2}), \ldots$ (with $t_{-i} = t_0 - ih$), one can similarly write $y'(t) \approx (y(t) - y(t-h))/h$ and derive the approximation

$$y_{-i-1} = y_{-i} - hf(t_{-i}, y_{-i})$$

for $i = 0, 1, 2, \ldots$

7.2. **An Example of Euler's Method.** Consider the ODE

$$y' = Ay, \quad y(t_0) = y_0$$

whose exact solution is $y(t) = e^{A(t-t_0)}y_0$. So imagine we fix real $t_{\text{end}} > t_0$, and use Euler's method to estimate $y(t_{\text{end}})$. For simplicity, we fix an (ideally) large integer $N > 0$, and set $h = (t_{\text{end}} - t_0)/N$, so that $t_N = t_{\text{end}}$ for this choice of h. Euler's method gives

$$y_{i+1} = y_i + hAy_i = (1 + Ah)y_i,$$

and hence, by induction, $y_i = (1 + Ah)^i y_0$ for all $i \geq 1$; hence the approximation of $y(t_{\text{end}})$ is

$$y_N = (1 + Ah)^N y_0 = \big(1 + A(t_{\text{end}} - t_0)/N\big)^N y_0.$$

There are a number of ways to derive the fact that for any $x \in \mathbb{R}$,

$$\lim_{N \to \infty} (1 + x/N)^N = e^x,$$

and therefore

$$\big(1 + A(t_{\text{end}} - t_0)/N\big)^N \xrightarrow{N\to\infty} e^{A(t_{\text{end}}-t_0)}.$$

Hence, as $N \to \infty$, Euler's formula approximation to $y(t_{\text{end}})$ tends to its true value.

7.3. **The (Explicit) Trapezoidal Method.** The (explicit) trapezoidal method often gives a better approximation than Euler's method. Given y_i an approximation for $y(t_i)$, it determines an approximation for $y(t_{i+1})$ by setting

$$Y_{i+1} = y_i + hf(t_i, y_i),$$

and then taking

$$y_{i+1} = y_i + h\frac{f(t_i, y_i) + f(t_{i+1}, Y_{i+1})}{2}.$$

Hence Y_{i+1} above looks like Euler's approximation, but it sets y_{i+1} using y_i and Y_{i+1}.

In class we explained: if $f(t, y) = f(t)$, then the solution to the initial value problem (8) is

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

In this case, Euler's method is like the "rectangular approximation," i.e., writing

$$y(t_0 + h) \approx y_0 + hf(t_0),$$

hence equivalent to

$$y_0 + \int_{s=t_0}^{s=t_0+h} f(s)\, ds \approx y_0 + hf(t_0),$$

i.e.,

$$\int_{s=t_0}^{s=t_0+h} f(s)\, ds \approx hf(t_0),$$

which is the usual "rectangle rule" for approximating the integral. However, for the Trapezoidal method we set

$$y(t_0 + h) \approx y(t_0) + h\frac{f(t_0) + f(t_1)}{2},$$

which is the usual "trapezoid rule" for approximating an integral.

In [A&G], it is explained that for fixed t_0, y_0, f and small h, we have

$$y(t_1) = y(t_0 + h) = y(t_0) + hf(t_0, y_0) + O(h^2).$$

By contrast, the trapezoidal method gives an $O(h^3)$ error, which typically an improvement (or no worse, even if f is not differentiable). For more details, CONTINUE HERE OR ASSIGN THE DETAILS FOR HOMEWORK.

POSSIBLE HOMEWORK PROBLEM: How does $y' = Ay$ perform under the (explicit) trapezoidal method? And: show that it converges faster as $N \to \infty$ in comparison to Subsection 7.2.

8. Similarity and a More Systematic Way of Computing e^{At} and A^n for Matrices A

To compute functions of a square matrix, A, such as e^{At} and A^n for $t \in \mathbb{R}$ and $n \in \mathbb{Z}$, we use *similarity*. The idea is that if we can find a matrices S, B such that S is invertible and

$$A = SBS^{-1},$$

then we have

$$A^2 = SBS^{-1}SBS^{-1} = SB^2S^{-1},$$

and similarly $A^k = SB^kS^{-1}$, and hence $f(A) = Sf(B)S^{-1}$ for any polynomial $f = f(x)$, and hence any globally convergent power series, f, and similarly any real analytic f, etc.

Remark 8.1. In class in 2024, we pointed out that if $A_i = SB_iS^{-1}$ for $i = 1, 2, 3$, then $A_3A_2A_1 = SB_3B_2B_1S^{-1}$. This was demonstrated by the instructor, taking S^{-1} to be "Joel moves from the podium to the blackboard," and B_1, B_2, B_3 to be, respectively, writing "Goo," "d af," "ternoon." on the blackboard (or something like that). We demonstrated that

$$(SB_3S^{-1})(SB_2S^{-1})(SB_1S^{-1})$$

indeed had the same effect as (the considerably more efficient) $SB_3B_2B_1S^{-1}$.

Next, given any A, we try to find a B that is as simple as possible such that $A = SBS^{-1}$. For example, if $A \in \mathbb{R}^{2\times 2}$ (i.e., A is a real 2×2 matrix), and

$$B = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix},$$

(for some $\lambda_1, \lambda_2 \in \mathbb{R}$, or $\lambda_1, \lambda_2 \in \mathbb{C}$) then we easily see that

$$B^2 = \begin{bmatrix} \lambda_1^2 & 0 \\ 0 & \lambda_2^2 \end{bmatrix},$$

and we similarly see that

$$f(B) = \begin{bmatrix} f(\lambda_1) & 0 \\ 0 & f(\lambda_2) \end{bmatrix}$$

where $f = f(x)$ is any polynomial, globally convergent power series, etc. It then follows that

$$f(A) = Sf(B)S^{-1} = S \begin{bmatrix} f(\lambda_1) & 0 \\ 0 & f(\lambda_2) \end{bmatrix} S^{-1}.$$

Definition 8.2. Let A be a $n \times n$ square matrix (over $\mathbb{R}$ or $\mathbb{C}$). We say that A is *diagonalizable* if for some $n \times n$ matrix S with complex entries, and real or complex numbers $\lambda_1, \ldots, \lambda_n$, we have

$$A = S \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_n \end{bmatrix}. \tag{6}$$

If $\lambda \in \mathbb{R}$ (or $\mathbb{C}$) and $A\mathbf{v} = \lambda\mathbf{v}$ for some vector $\mathbf{v} \neq \mathbf{0}$ (so $\mathbf{v} \in \mathbb{R}^n$ or $\mathbb{C}^n$), we say that λ is an *eigenvalue of A* and any such $\mathbf{v}$ is an *eigenvector of A corresponding to the eigenvalue λ*; we also call $(\lambda, \mathbf{v})$ an *eigenpair* of A.

Remark 8.3. If (6) holds, then the columns of S are linearly independent, and the i-th column, $\mathbf{v}_i$, of S forms an eigenpair $(\lambda_i, \mathbf{v}_i)$ (i.e., $A\mathbf{v}_i = \lambda_i \mathbf{v}_i$); the converse also holds.

In the next section we will see that not all matrices are diagonalizable.

In class we showed examples that to illustrate that if a 2×2 matrix has eigenpairs $(\lambda_1, \mathbf{v}_1)$ and $(\lambda_2, \mathbf{v}_2)$, and if S has its columns consisting of $\mathbf{v}_1$ and $\mathbf{v}_2$, then

$$AS = SB, \quad \text{where} \quad \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix},$$

since multiplying on the right by B operates on the columns of S. If $\mathbf{v}_1, \mathbf{v}_2$ are linearly independent, and then S are invertible as well, and hence

$$A = SBS^{-1}$$

is a "diagonalization of A."

One way to find eigenpairs is to note that if $A\mathbf{v} = \lambda v = \lambda I \mathbf{v}$, then $(A - \lambda I)\mathbf{v} = \mathbf{0}$. Hence, in this case we can say that $M = A - \lambda I$ has any of the equivalent properties: (1) M has a non-trivial nullspace, (2) M is not invertible, (3) M has rank less than n, (4) $\det(M) = 0$, (5) the row echelon form of M has a row of 0's, (6) the image of M is of dimension less than n, (7) etc. One method for finding eigenvalues of a 2×2 matrix is to solve the equation $\det(A - \lambda I) = 0$; since $\det(A - \lambda I)$ is a polynomial with leading term λ^2, there are always two solutions (at least over $\mathbb{C}$).

We remark that if A is $n \times n$ with n odd, then $\det(A - \lambda I)$ is a polynomial with leading term $-\lambda^n$; for this (and other) reasons, one tends to work with $\det(\lambda I - A)$ instead (which is called the *characteristic polynomial of* A).

[Looking for solutions of $\det(\lambda I - A) = 0$ when A is a large square matrix tends to be numerically impractical; there are a lot of other general methods to find eigenpairs, especially in special types of matrices.]

Example 8.4. Say that

$$A = \begin{bmatrix} 2 & 4 \\ 3 & 6 \end{bmatrix}.$$

Then

$$\det(\lambda I - A) = \det \begin{bmatrix} \lambda - 2 & -4 \\ -3 & \lambda - 6 \end{bmatrix} = (\lambda - 2)(\lambda - 6) - 12 = \lambda^2 - 8\lambda.$$

Hence A has two eigenvalues, $\lambda = 0, 8$. To find a corresponding eigenvector to $\lambda = 8$, we solve for

$$(8I - A)\mathbf{v} = \mathbf{0},$$

hence

$$\begin{bmatrix} 6 & -4 \\ -3 & 2 \end{bmatrix} \mathbf{v} = \mathbf{0},$$

and we find that any (non-zero) multiple of $\mathbf{v}_1 = [2/3 \ 1]$ yields a corresponding eigenvector. Similarly for $\mathbf{v}_2 = [-2 \ 1]$. Hence

$$A = S \begin{bmatrix} 8 & 0 \\ 0 & 0 \end{bmatrix} S^{-1}, \quad \text{where} \quad S = \begin{bmatrix} 2/3 & -2 \\ 1 & 1 \end{bmatrix}.$$

In particular

$$e^{At} = S \begin{bmatrix} e^{8t} & 0 \\ 0 & 1 \end{bmatrix} S^{-1},$$

(since $e^{0 \cdot t} = 1$), and similarly for A^n.

There are a number of families of matrices with well-known eigenvectors and/or eigenvalues. For example, for any $a, b \in \mathbb{R}$,

$$A = \begin{bmatrix} a & b \\ b & a \end{bmatrix}$$

has eigenpairs $(a+b, [1\ 1])$ and $(a-b, [1\ -1])$; this is an example of (square) Toeplitz matrix, whose eigenvectors in the $n \times n$ case are known to be $\mathbf{v}_\zeta = [1\ \zeta\ \zeta^2\ \cdots\ \zeta^{n-1}]$ where $\zeta^n = 1$, and corresponding eigenvalues are given by the "(literal, not complex) dot product" of ζv with the top row of the matrix).

As another example, if

$$A = \begin{bmatrix} a & b \\ -b & -a \end{bmatrix}$$

then the eigenvalues are $\pm\sqrt{a^2 - b^2}$, which are necessarily complex if $|b| > |a|^2$.

8.1. **Eigenvalues of Other Matrices.** One can get a lot of important intuition about eigenvalues and their meaning by consider other classes of matrices.

Example 8.5. Let

$$A = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix},$$

which is counterclockwise rotation by θ (i.e., addition by $+\theta$ in the angular polar coordinate) in $\mathbb{R}^2$ (i.e., $\mathbf{v} \mapsto A\mathbf{v}$ takes $(1, 0)$ to $(\cos\theta, \sin\theta)$ and $(0, 1)$ to $(-\sin\theta, \cos\theta)$. Then A has eigenvalues $\lambda = e^{\pm i\theta}$ where $i = \sqrt{-1}$. Hence A^n, after a similarity transformation, looks like a diagonal matrix with diagonal entries $e^{\pm i\theta n}$.

Example 8.6. Let $V \subset \mathbb{R}^n$ be a d-dimensional subspace of $\mathbb{R}^n$, and let $P\colon \mathbb{R}^n \to V$ be the orthogonal projection to V. Then if V' is the $n-d$ orthogonal complement of V, P takes V' to $\mathbf{0}$, and is the identity map on V. It follows that P has eigenvalues 0 and 1, with 0 having multiplicity $n - d$ and 1 having multiplicity d.

9. Finite Recurrences and Relationship to ODE's

One of the most famous recurrences (with constant coefficients) is the *Fibonacci recurrence equation*

$$F_{n+2} = F_{n+1} + F_n, \quad \text{subject to } F_1 = 1,\ F_2 = 1,$$

which yields the Fibonacci numbers

$$F_3 = 2,\ F_4 = 3,\ F_5 = 5,\ F_6 = 8,\ F_7 = 13,\ F_8 = 21,\ F_9 = 34, \ldots$$

and, writing $F_n = F_{n+2} - F_{n+1}$, and setting $n = 0, -1, -2, \ldots$ we get

$$F_0 = 0,\ F_{-1} = 1,\ F_{-2} = -1,\ F_{-3} = 2,\ F_{-4} = 3,\ F_{-5} = -5\ F_{-6} = 8,\ F_{-7} = -13, \ldots$$

We study finite recurrences for a few reasons: (1) they are the discrete analog of ODE's; (2) ODE approximation methods (e.g., Euler's method, explicity trapezoidal rule) *are actually finite recurrence equations*; and (3) it is easier to see what goes wrong with finite recurrences that the analogous ODE solvers. For an example of (3), it is easy to see that the Fibonacci recurrence $F_{n+2} = F_{n+1} + F_n$ is problematic as $n \to \infty$ for a solution $F_0 = 1$ and $F_1 = (1 - \sqrt{5})/2$; you can see this by typing in MATLAB:

```
clear
F{1}=1
F{2}=(1-sqrt(5))/2
for i=3:1000, F{i}=F{i-1}+F{i-2}; end
```

If you examine `F{20}` and `F{2}^19`, things look pretty good. (just try typing in differentiation, it is sometimes easier to see what goes wrong with finite recurrences. If you type `for i=1:10:141, {i,F{2}^(i-1),F{i}}, end`, you start to see the effect of roundoff/truncation. And if you examine `F{1000}` and `F{2}^999`, things look rediculous.

In terms of ODE's, it best to view the Fibonacci numbers as defined by a *finite recurrence (equation)*

$$F_{n+2} - F_{n+1} - F_n = 0, \quad \forall n \in \mathbb{Z}, \tag{7}$$

and *initial conditions*

$$F_1 = 1,\ F_2 = 1.$$

It is immediate this equation plus initial conditions determine F_n for integers $n \geq 3$; since one can also solve for F_n as $F_{n+2} - F_{n+1}$, these initial conditions also uniquely determine F_n for integers $n \leq 0$.

Just as for second order ODE's, one can write

$$\mathbf{y}_n = \begin{bmatrix} F_{n+1} \\ F_n \end{bmatrix}$$

and write (7) as a two-term recurrence

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

and hence

$$\mathbf{y}_{n+1} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \mathbf{y}_n.$$

Hence, by induction we have

$$\mathbf{y}_n = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}^n \mathbf{y}_0.$$

It turns out (just as we guessed that $\mathbf{y}' = A\mathbf{y}$ has $\mathbf{y}(t) = e^{At}\mathbf{C}$ as its general solution), there is a standard way to "guess" solutions to finite recurrence equations. To solve (7), we "guess" that $F_n = r^n$ might work for certain $r \in \mathbb{C}$, and we see that (7) holds iff

$$r^{n+2} = r^{n+1} + r^n \quad \forall n \in \mathbb{Z}.$$

Assuming that $r \neq 0$, this is equivalent to

$$r^2 = r + 1,$$

whose solution is the golden ratio and its "conjugate"[4]

$$r = \frac{1 \pm \sqrt{5}}{2}.$$

[4] Here "conjugate" is actually a precise term in algebra: it is the conjugate over the degree two extension $\mathbb{Q}(\sqrt{5})$ over $\mathbb{Q}$.

Notice that (7) is a "linear recurrence," meaning that any linear combination of solutions to (7) is again a solution; it follows that

$$F_n = c_1\left(\frac{1+\sqrt{5}}{2}\right)^n + c_2\left(\frac{1-\sqrt{5}}{2}\right)^n.$$

Since for any given values of F_0, F_1, one can find c_1, c_2 such that

$$c_1\left(\frac{1+\sqrt{5}}{2}\right)^0 + c_2\left(\frac{1-\sqrt{5}}{2}\right)^0 = F_0$$

$$c_1\left(\frac{1+\sqrt{5}}{2}\right)^1 + c_2\left(\frac{1-\sqrt{5}}{2}\right)^1 = F_1$$

it follows that all solutions to (7) can be written in this way. We remark that

$$\begin{bmatrix} \left(\frac{1+\sqrt{5}}{2}\right)^0 & \left(\frac{1-\sqrt{5}}{2}\right)^0 \\ \left(\frac{1+\sqrt{5}}{2}\right)^1 & \left(\frac{1-\sqrt{5}}{2}\right)^1 \end{bmatrix}$$

is an example of a 2×2 Vandermonde matrix. In general, the solution of any recurrence relation with constant coefficients will produce a *generalized Vandermonde matrix*; we'll say more about this when we cover interpolation and when we cover more about numerical differentiation and Taylor's theorem.

For more examples regarding recurrences, the reader can consult "CPSC 303: Recurrence Relations and Finite Recurrences." This article was created in version of the course that did not contain a three-week (or so) introduction to ODE's and MATLAB; hence this article is self-contained and motivates recurrences as a way to test the limits of double precision.

10. More on ODE's Arising in Celestial Mechanics and Newton's Gravitational Law

---

Appendix A. Fundamental Theorems Regarding ODE's

In this section we consider an *ODE initial value problem*, by which we mean an equation

$$\mathbf{y}'(t) = \mathbf{f}\big(t, \mathbf{y}(t)\big), \quad \mathbf{y}(t_0) = \mathbf{y}_0, \tag{8}$$

or more concisely

$$\mathbf{y}' = \mathbf{f}(t, \mathbf{y}), \quad \mathbf{y}(t_0) = \mathbf{y}_0, \tag{9}$$

where $\mathbf{f}$ is defined near $(t_0, \mathbf{y}_0)$, i.e., specifically on a "closed neighbourhood" of $(t_0, \mathbf{y}_0)$ of the form:

$$\mathcal{N}_{\delta_1,\delta_2} = \big\{(t, \mathbf{y}) \;\big|\; |t - t_0| \le \delta_1, \|\mathbf{y} - \mathbf{y}_0\| \le \delta_2\big\}$$

for some $\delta_1, \delta_2 > 0$; hence $\mathbf{f}\colon \mathcal{N}_{\delta_1,\delta_2} \to \mathbb{R}^m$. See [A&G], Section 16.1, page 482, which uses $a, \mathbf{c}$ instead of $t_0, \mathbf{y}_0$. Often $\mathbf{f}$ extends to a function $\mathbb{R}\times\mathbb{R}^m \to \mathbb{R}$. By a *local solution* of (8) or (9), we mean a differentiable function

$$y\colon (t_0 - \epsilon, t_0 + \epsilon) \to \mathbb{R}^m$$

satisfying

$$\mathbf{y}(t_0)=\mathbf{y}_0, \qquad \text{and} \qquad \forall |t-t_0|<\epsilon, \quad \mathbf{y}'(t)=\mathbf{f}\big(t,\mathbf{y}(t)\big). \tag{10}$$

A.1. **The Existence Theorem.**

Theorem A.1. *Consider an initial value problem (9) where* $\mathbf{f}$ *is a continuous function defined in a neighbourhood of* $(t_0,\mathbf{y}_0)$*. Then there exists a local solution (10) to (9). More precisely, let* $\|\cdot\|$ *be any norm on* $\mathbb{R}^m$*, and say that for* $\delta_1,\delta_2>0$*,* $\mathbf{f}\colon \mathcal{N}_{\delta_1,\delta_2}\to\mathbb{R}^m$ *is continuous and that* $\|\mathbf{f}\|\le B$ *in all of* $\mathcal{N}_{\delta_1,\delta_2}$*. Then in (10) we may take* $\epsilon=\max(\delta_1,\delta_2/B)$*, and* $\mathbf{y}$ *satisfies*

$$\|\mathbf{y}(t)-\mathbf{y}(s)\|\le B|t-s|$$

for all $s,t\in(t_0-\epsilon,t_0+\epsilon)$*.*

The usual way to conceptualize the proof below is by stating the Arzelà-Ascoli Lemma as an intermediate step. Here we give the entire proof from scratch.

Proof. We use Euler's method, run forwards and backwards: fix $f,t_0,\mathbf{y}_0,\delta_1,\delta_2,B,\epsilon$ as in the theorem statement. For each $h>0$ define $f_h(t)$ as follows: let N be the largest natural number such that $Nh<\epsilon$; for each integer i with $|i|\le N$, let $t_i=t_0+ih$. Hence for all $0\le i\le N$ we have

$$|t_i-t_0|\le h|i|\le hN<\epsilon<\delta_1. \tag{11}$$

Next, for each $i=0,1,2,\dots,N-1$, we inductively define $\mathbf{y}_{i+1}$ and $\mathbf{y}_{-(i+1)}$ via

$$\mathbf{y}_{i+1}=\mathbf{y}_i+h\mathbf{f}(t_i,\mathbf{y}_i),$$

and

$$\mathbf{y}_{-(i+1)}=\mathbf{y}_{-i}-h\mathbf{f}(t_{-i},\mathbf{y}_{-i}).$$

As we do so, by induction on $i=0,\dots,N-1$ we claim that

$$\|\mathbf{y}_{i+1}-\mathbf{y}_i\|\le Bh$$
$$\|\mathbf{y}_{-(i+1)}-\mathbf{y}_{-i}\|\le Bh;$$

to see this, note that (1) these inequalities hold for $i=0$ since $(t_0,\mathbf{y}_0)\in\mathcal{N}_{\delta_1,\delta_2}$ and hence; $\|\mathbf{f}(t_0,\mathbf{y}_0)\|\le B$, and (2) if these inequalities hold with i replaced by any integer less than i, then

$$\|\mathbf{y}_i-\mathbf{y}_0\|\le\|\mathbf{y}_i-\mathbf{y}_{i-1}\|+\cdots+\|\mathbf{y}_1-\mathbf{y}_0\|\le Bih\le BNh<\delta_2$$

and similarly for $\|\mathbf{y}_{-i}-\mathbf{y}_0\|$; hence $(t_i,\mathbf{y}_i),(t_{-i},\mathbf{y}_{-i})$ both lie in $\mathcal{N}_{\delta_1,\delta_2}$. Hence,

$$\|\mathbf{y}_{i+1}-\mathbf{y}_i\|\le Bh.$$

and similarly for $\|\mathbf{y}_{-(i+1)}-\mathbf{y}_{-i}\|$.

Now we define $y_h\colon(t_0-\epsilon,t_0+\epsilon)\to\mathbb{R}^m$ by setting

$$\forall t_0\le t<t_0+\epsilon, \quad y_h(t)=y_i \quad \text{where } i=\lfloor(t-t_0)/h\rfloor,$$

and

$$\forall t_0-\epsilon<t\le t_0, \quad y_h(t)=y_i \quad \text{where } i=-\lfloor-(t-t_0)/h\rfloor.$$

From the above we have that y_h is defined for all t with $|t-t_0|<\epsilon$. Moreover, we claim that for any $t',t''\in(t_0-\epsilon,t_0+\epsilon)$ we have

$$\|\mathbf{y}_h(t')-\mathbf{y}_h(t'')\|\le B(|t'-t''|+2h), \tag{12}$$

since we may assume $t' < t''$, and then take the closest $t_{i'} = t_0 + hi'$ to t' with $t_{i'} \geq t'$, and similarly define $t_{i''}$, and then estimate the above left-hand-side of (12) via

$$\|\mathbf{y}_h(t')-\mathbf{y}_h(t'')\| \leq \|\mathbf{y}_h(t')-\mathbf{y}_h(t_{i'})\|+\|\mathbf{y}_h(t_{i'})-\mathbf{y}_h(t_{i''})\|+\|\mathbf{y}_h(t_{i''})-\mathbf{y}_h(t'')\| \leq h+B|t_{i'}-t_{i''}|+h,$$

which establishes (12).

So consider the sequence $y_1, y_{1/2}, y_{1/3}, y_{1/4}, \ldots$. From (12) we have that $y_{1/n}(t)$ is bounded for each $t \in (t_0-\epsilon, t+\epsilon)$. Since the set of rational numbers in $(t_0-\epsilon, t+\epsilon)$ is countable, by passing to successive subsequences and taking a diagonal subsequence, we can produce a sequence of naturals $n_1 < n_2 < \ldots$ such that for any rational $r \in (t_0 - \epsilon, t + \epsilon)$, we have that $y_{1/n_i}(r)$ converges. So define $f(r)$ for each such rational as the limit of $y_{1/n_i}(r)$.

Hence y is defined on all rational numbers in the interval $I = (t_0 - \epsilon, t + \epsilon)$. Let us show that $y(r)$ extends to a unique function y defined on the entire interval.

It follows that for any $t', t'' \in (t_0 - \epsilon, t + \epsilon)$ and integers $i' < i''$, (12) implies that

$$\|\mathbf{y}_{1/n_{i'}(t')} - \mathbf{y}_{1/n_{i''}}(t'')\| \leq B(|t' - t''| + 2/n_{i'}). \tag{13}$$

Hence for any rational $r', r'' \in I$ we have, taking $i' \to \infty$,

$$\|\mathbf{y}(r') - \mathbf{y}(r'')\| \leq B|r' - r''|. \tag{14}$$

It follows that if $t \in I$, and $r_1, r_2, \ldots$ is any sequence of rationals tending to t, $\mathbf{y}(r_i)$ is a Cauchy sequence, and therefore has a limit $\mathbf{y}(t)$; (14) shows that this limit is independent of the choice of the sequence, and that for any $t' \in I$, by taking a sequence of rationals tending to t, and another to t', an argument similar to establishing (12) establishes

$$\|\mathbf{y}(t) - \mathbf{y}(t')\| \leq B|t - t'|. \tag{15}$$

We also remark that for all $t \in I$, we have $(t, \mathbf{y}(t)) \in \mathcal{N}$, and hence

$$g(x) \stackrel{\text{def}}{=} \max_{t,t' \in I \text{ and } |t-t'|=x,} \|\mathbf{f}(t, \mathbf{y}(t) - \mathbf{f}(t', \mathbf{y}(t'))\| \tag{16}$$

exists (since $\mathcal{N}$ is closed), and moreover $g(x) \to 0$ as $x \to 0$ (if not, we take a sequence of x_i, t_i, t_i' with $x_i \to 0$, $|t_i - t_i'| = x_i$, and the above right-hand-side, with t_i, t_i' replacing t, t', bounded away from 0; this contradicts the closure of $\mathcal{N}$ and the continuity of f). Now let us prove that $\mathbf{y}(t)$ is a solution to the initial value problem in question.

The estimate (15) and (16) shows that for any $t \in I$ we have

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

exists, and equals the limit of its Riemann sums starting from either endpoint (note that for $n \geq 2$, this Riemann sum is really a sum of vectors, not reals). The fact that for any i and $t \in I$ with $t > 0$ we have

$$\mathbf{y}_{1/n_i}(t) = \mathbf{y}_0 + (1/n_i)\Big(\mathbf{f}\big(1/n_i, \mathbf{y}_{1/n_i}(1/n_i)\big) + \cdots + \mathbf{f}_{1/n_i}\big(1/n_i, \mathbf{y}_{1/n_i}(\lfloor tn_i \rfloor/n_i)\big)\Big);$$

note also that

$$\left|\int_{s=t_0}^{s=t} \mathbf{f}\big(s, \mathbf{y}(s)\big)\, ds - (1/n_i)\Big(\mathbf{f}\big(1/n_i, \mathbf{y}(1/n_i)\big) + \cdots + \mathbf{f}\big(1/n_i, \mathbf{y}(\lfloor tn_i \rfloor/n_i)\big)\Big)\right| \leq (1/n_i)\lfloor tn_i \rfloor \max_{0 \leq x \leq 1/n_i} g(x) \leq G(1/n_i)$$

where

$$G(x) \stackrel{\text{def}}{=} \max_{0\le x'\le x} g(x),$$

and g is as in (16). Taking $i\to\infty$ we get

$$\left|\mathbf{y}(t) - y_0 - \int_{s=t_0}^{s=t} \mathbf{f}\big(s,\mathbf{y}(s)\big)\,ds\right| \le \lim_{i\to\infty} G(1/n_i) = 0.$$

Hence $\mathbf{y}(t)$ is a solution to the ODE in question, for all $t\in I$. □

Remark A.2. The example $y' = |y|^{1/2}$ shows that you really have to define $\mathbf{y}$ above by taking a subsequence $\mathbf{y}_{1/n_i}$ that converges on all rationals, I, or at a subset thereof whose closure contains the endpoint of I: indeed, in the above proof you can take δ_1 arbitrarily large, and with $(t_0,y_0) = (0,0)$ you can take $\delta_2 = C$ and $B = \sqrt{C}$ for any real $r > 0$. Then $I = (-\sqrt{C},\sqrt{C}) = (-B,B)$; if y_{1/n_i} doesn't converge near, say, $\sqrt{B}$, then there are different possible limiting values for y_{1/n_i} near $\sqrt{B}$, and by oscillating between any two such solutions to this ODE, you cannot hope to have the limit $y_{1/n_i}(\sqrt{B})$ to exist.

Appendix B. Preliminary Facts from Advanced Calculus

Many readers will know these lemmas; others who don't may wish to skip the proofs. The proofs are standard and comprise some basic tools of advanced calculus.

Lemma B.1. *There is an infinite sequence $q_1,q_2,\dots$ of rational numbers that contains each rational number at least once.*

Proof. We write the rationals in "phases," namely

Phase 1: 0
Phase 2: $1/1, -1/1$
Phase 3: $2/1, -2/1, 1/2, -1/2,$
Phase 4: $3/1, -3/1, 2/2, -2/2, 1/3, -1/3,$

and so on; hence each "phase" has finitely many rationals, and the fraction $\pm a/b$ with $a,b\in\mathbb{N}$ appears in phase $a+b$. Combining the phases starting with Phase 1 and onward, we get the sequence

$$0, 1/1, -1/1, 2/1, -2/1, 1/2, -1/2, 3/1, -3/1, 2/2, -2/2, 1/3, -1/3, \dots$$

which contains every rational at least once. □

Lemma B.2. *Let $r_1,r_2,\dots$ be a sequence of bounded real numbers, i.e., $|r_i|\le M$ for some $M\in\mathbb{R}$. Then the sequence has a convergent subsequence, i.e., there are integers $n_1 < n_2 < n_3 < \dots$ such that*

$$\lim_{i\to\infty} r_{n_i} = r$$

for some real r (with $|r|\le M$).

Proof. Introduce the notation $n_{1,i} = i$, and let $I_1 = [-M,M]$, which is a closed interval of length $2M$.

Since $I_1 = [-M,0]\cup[0,M]$, there are either infinitely many r_i in $[-M.0]$, or infinitely many in $[0,M]$ (or infinitely many in both). Let I_2 contain infinitely many, with I_2 being either $[-M,0]$ or $[M,0]$; hence I_2 is of length $M = |I_1|/2$,

and for some $n_{2,1} < n_{2,2} < n_{2,3} < \dots$ we have $r_{n_{2,i}} \in I_2$ for all i. Hence we have produced sequences

$$\begin{array}{cccc} n_{11} = 1, & n_{12} = 2, & n_{13} = 3, & \dots \\ n_{21}, & n_{22}, & n_{23}, & \dots \end{array}$$

such that the lower sequence is a subsequence of the upper sequence.

Similarly, we can find an interval I_3 of length $|I_2|/2 = |I_1|/4$, and a subsequence $n_{31}, n_{32}, \dots$ of $n_{21}, n_{22}, n_{23}, \dots$ such that $r_{3i} \in I_3$ for all i. Continuing in this fashion we get an array

$$\begin{array}{cccc} n_{11} = 1, & n_{12} = 2, & n_{13} = 3, & \dots \\ n_{21}, & n_{22}, & n_{23}, & \dots \\ n_{31}, & n_{32}, & n_{33}, & \dots \\ \vdots, & \vdots, & \vdots, & \ddots \end{array}$$

of successive subsequences such that for each $j \in \mathbb{N}$, $n_{j1}, n_{j2}, \dots \in I_j$, where I_j is an interval of size $M/2^{j-1}$.

We now claim that the "diagonal sequence" $r_{n_{1,1}}, r_{n_{2,2}}, r_{n_{3,3}}, \dots$ has a limit. The details are omitted. □

Exercises

(1) Here will appear Exercise 1.
(2) Here will appear Exercise 2.
(3) Here will appear Exercise 3.
(4) Etc.

Department of Computer Science, University of British Columbia, Vancouver, BC V6T 1Z4, CANADA.

E-mail address: jf@cs.ubc.ca
URL: http://www.cs.ubc.ca/~jf