Applied Differential Calculus

LECTURE 1: First-order ordinary differential equations.

Authors:

Manuel Carretero, Luis L. Bonilla, Filippo Terragni, Sergei Iakunin, Rocío Vega

Departamento de Ciencia e Ingeniería de Materiales e Ingeniería Química,

Universidad Carlos III de Madrid,

Bachelor's Degree in Computer Science and Engineering and

Dual Bachelor in Computer Science and Engineering and Business Administration.

I. BASICS

1. **Differential equation** is a relation between a function (*the unknown*) and its derivatives. **To solve the differential equation** means finding the function or family of functions that satisfy the equation.

2. If the unknown is a function of one variable, the differential equation is called an **ordinary differential equation (ODE)**. If it is a function of several variables, we have a partial differential equation (PDE).

3. **Order of a differential equation** is the order of the highest derivative appearing in it.

4. If the unknown is a vector function and so is the relation between its derivatives and the unknown, we have a **system of (ordinary or partial) differential equations**.

5. **Linear ordinary differential equation of order** N:
   These equations can be written in the form
   $$\sum_{n=0}^{N} a_n(x)\, u^{(n)}(x) = F(x). \tag{1}$$
   Where the derivatives of the unknown function $u^{(0)}(x) = u(x)$ are $u^{(n)}(x) = d^n u/dx^n$, $n = 1, 2 \cdots N$.
   On the other hand, $a_0(x), a_1(x), \cdots, a_N(x)$ and $F(x)$ are known functions.
   Any other ODE is ***nonlinear***.
   If $F(x) \equiv 0$, (1) is a **homogeneous linear ODE**; if not, it is an **inhomogeneous linear ODE**. $F(x)$ is often called the **source term**.

6. An Nth order ODE can be written as a **system of** N **first order ODEs**.
   For example, (1) is equivalent to
   $$u_0' = u_1, \quad \ldots, \quad u_{n-1}' = u_n, \quad \ldots, \quad u_{N-1}' = \frac{F(x) - \sum_{n=0}^{N-1} a_n(x)\, u_n}{a_N(x)},$$
   where $u_0(x) = u^{(0)}(x) \equiv u(x)$, $u_n(x) = u^{(n)}(x)$, $n = 1, 2, \cdots, N-1$.

7. There are differential equations whose solutions can be found explicitly by **analytical methods**. These equations are a small class of all possible differential equations and,

in many cases, they are linear. In some cases, the differential equation we want to solve is **close** to another one whose solution we can find analytically. There are techniques to exploit this 'closeness' to find approximations to the solutions of the equation of interest. You will find examples of this line of reasoning in other engineering, physics or applied mathematics courses. In many cases, we can find **approximate solutions** of the differential equations and additional **boundary conditions** by **numerical methods** that replace the original differential equation by difference equations or by algebraic problems that are solved with a computer.

We will see several numerical methods for ODEs and PDEs in this course.

Examples.

Differential equations appear naturally because we try to understand phenomena by seeking relations between rates of variation of magnitudes and the magnitudes themselves according to physical reasoning. Consider for instance, a large number of cows in an enclosed valley with abundant pasture. Let $u(t)$ be the **number of cows per hectare**. A simple assumption about the growth of cow population is that their growth, $\dot{u} = du/dt$ is proportional to u:

$$\frac{du}{dt} = r\,u. \tag{2}$$

The proportionality constant $r > 0$ is called the **birth rate**.

Equation (2) is a first order linear homogeneous ODE of the form (1) with $a_1 = 1$, $a_0 = -r$, $F = 0$.

A typical problem (called an **initial value problem (IPV)**) is to find $u(t)$ knowing the cow density $u(0) = u_0$ at a given time $t = 0$.

Solution by ***separation of variables:***

$$\int^u \frac{du}{u} = \int^t r\,dt \Longrightarrow \ln u = rt + \mu \Longrightarrow u(t) = c\,e^{rt}, \tag{3}$$

where $c = e^{\mu}$ is an arbitrary constant.

We can show that (3) is the most general solution of the ODE (2): write $u(t) = e^{rt}v(t)$ and substitute in the ODE. We get $dv/dt = 0$ which immediately implies $v = c$ (constant).

Of all these solutions, only one obeys the initial condition (IC): $u(0) = c$ implies $c = u_0$, which is the only value of c that satisfy the IC.

Note that the two relations, the ODE (3) and the initial condition $u(0) = u_0$ ***uniquely determine the density*** $u(t)$***.***

This is typical of mathematical models arising from the sciences: they give rise to differential equations with a number of boundary conditions whose solution is unique. Mathematical conditions that guarantee existence and uniqueness of the solution to initial value problems will be explained later.

Physical meaning of the solution:

The initial cow population u_0 is doubled at time $t_1 = \ln 2/r$ and it becomes $2^n u_0$ at time $t_n = n \ln 2/r$.

This population explosion applied to the human population on earth is known as **Malthus law** (1798).

For example, assuming that the population grows 2% per year, the birth rate is $r = 0.02/\text{year}$ and the population will double at time $t_1 = 50 \ln 2 \approx 34.65$ years and quadruplicate in about 70 years.

Of course an enclosed valley has limited resources and may correct (2) by imagining that the birth rate becomes zero when the carrying capacity of the valley (maximum cow density stably supported by the valley) C is reached.

Thus we would replace the constant r in the ODE by $r(1 - u/C)$ thereby obtaining the ***logistic*** **equation**

$$\frac{du}{dt} = r\,u\Big(1 - \frac{u}{C}\Big), \tag{4}$$

proposed by **F. Verhulst** (1838). (4) can also be solved by separation of variables although the algebra is now a little bit more involved.

The result is

$$u(t) = \frac{C}{Ce^{-rt-\mu} + 1} = \frac{Cu_0}{u_0 + (C - u_0)e^{-rt}}, \tag{5}$$

once the constant of integration μ is calculated so that $u(0) = u_0$. It is obvious that any nonzero initial cow density will evolve to the carrying capacity C as $t \to \infty$.

There are two **critical points** (also called **equilibrium solutions**) $u = 0$ and $u = C$ which are **constant solutions** of (4).

Clearly, initial conditions near $u = 0$ give rise to solutions that move away from it whereas initial conditions in a neighborhood of $u = C$ give rise to solutions that **approach** it.

The critical point $u = 0$ is ***unstable*** whereas $u = C$ is **asymptotically stable**.

II. FIRST-ORDER LINEAR ODE

We now turn to the solution of the general first-order linear ODE (1):

$$a_1(x)\, u' + a_0(x)u = F(x). \tag{6}$$

To find the general solution, we first solve the corresponding **homogeneous equation** with $F = 0$ by separation of variables.

The result is

$$u_h(x) = e^{-g(x)}, \quad g(x) = \int^x \frac{a_0(t)}{a_1(t)}\, dt. \tag{7}$$

We now multiply the inhomogeneous equation (6) by $e^{g(x)}$ and divide by $a_1(x)$.

We observe that the result may be written as

$$[e^{g(x)}u]' = \frac{F(x)e^{g(x)}}{a_1(x)}, \tag{8}$$

which is immediately integrated to produce

$$u(x) = e^{-g(x)} \int^x e^{g(t)} \frac{F(t)}{a_1(t)}\, dt = \int^x e^{-\int_t^x a_0(s)/a_1(s) ds} \frac{F(t)}{a_1(t)}\, dt. \tag{9}$$

$e^{g(x)}/a_1(x)$ is called ***an integrating factor*** for (6) because it reduces it to (8) that can be solved by **direct integration**.

Any two possible choices of $g(x)$ differ by a constant of integration (which we may denote by μ) which gets cancelled in (9).

Thus the choice of $g(x)$ does not matter.

However, selecting one **particular primitive** of

$$P(x) = e^{-g(x)} \int^x e^{g(t)} \frac{F(t)}{a_1(t)}\, dt, \tag{10}$$

the most **general solution** of (6) can be written as

$$u(x) = Ae^{-g(x)} + P(x), \tag{11}$$

where A is a constant.

The result (11) says that **the general solution** of the inhomogeneous linear ODE (6) is ***the sum of a particular solution*** $P(x)$ ***of the inhomogeneous linear ODE plus a general solution of the associated homogeneous equation***.

It turns out that this statement, known as the ***superposition principle*** is also true for the general Nth-order linear ODE (1):
If $P(x)$ is a particular solution of (1) and $u_j(x)$, $j = 1, \ldots, R$ are solutions of the associated homogeneous equation with $F = 0$, $P(x) + \sum_{j=1}^{R} A_j u_j(x)$ is also a solution of (1) for any numbers A_j.

Examples.

1. $u' + xu = xe^{x^2}$ with $u(0) = 1$.
   Observe that $e^{x^2/2}$ is an **integrating factor** (this is better done by inspection, not by the laborious procedure of using (7) unless you are beaten after trying for a while).
   Then $(e^{x^2/2}u)' = xe^{3x^2/2}$ from which $e^{x^2/2}u(x) = \frac{1}{3}e^{3x^2/2} + c$ and therefore $u(x) = \frac{1}{3}e^{x^2} + ce^{-x^2/2}$.
   The initial condition gives $1 = c + 1/3$ and thus $u(x) = \frac{1}{3}[e^{x^2} + 2e^{-x^2/2}]$.

2. **Variation of parameters** provides the same solution. $u_h = Ce^{-x^2/2}$ is the solution of the homogeneous equation. Replace the ***parameter*** C by a function $v(x)$ (***variation of parameters***) and substitute $u(x) = e^{-x^2/2}v(x)$ in the inhomogeneous ODE.
   The result is $u' + xu = v'e^{-x^2/2} = xe^{x^2}$.
   This yields the simpler ODE $v' = xe^{3x^2/2}$ which can be integrated immediately: $v = \frac{1}{3}e^{3x^2/2} + c$.
   The corresponding solution of the inhomogeneous ODE is $u(x) = \frac{1}{3}e^{x^2} + ce^{-x^2/2}$.

3. The result that the solution of the linear first-order inhomogeneous ODE is the sum of a particular solution of the ODE plus a general solution of the corresponding homogeneous ODE suggests **another solution method**.
   Try $u_p(x) = Ae^{x^2}$ as a **particular solution** of the inhomogeneous ODE $u' + xu = xe^{x^2}$ and calculate the **undetermined coefficient** A:
   $u_p' + xu_p = A2xe^{x^2} + xAe^{x^2} = 3Axe^{x^2}$ should be equal to xe^{x^2}. This gives $3A = 1$ or $A = 1/3$.
   Then the particular solution is $u_p(x) = \frac{1}{3}e^{x^2}$, the **general solution** of the homogeneous equation is $u_h(x) = ce^{-x^2/2}$ and the general solution is the sum $u(x) = \frac{1}{3}e^{x^2} + ce^{-x^2/2}$, as before.

4. $u' + xu = x^{2m+1}$ with $u(1) = 1$ and $m = 0, 1, \ldots$.

   We find $(e^{x^2/2}u)' = e^{x^2/2}x^{2m+1} = x^{2m}(e^{x^2/2})'$ whose right hand side can be found exactly by integration by parts:

   $I_m = \int e^{x^2/2}x^{2m+1}dx = x^{2m}e^{x^2/2} - 2mI_{m-1}$.

   The solution of the first-order linear difference equation $I_m + 2mI_{m-1} = f_m$ is $I_m = (-2)^m m! I_0 + \sum_{j=1}^{m}(-2)^{m-j} f_j m!/j!$ and therefore the general solution of the ODE is $u(x) = ce^{-x^2/2} + \sum_{j=0}^{m}(-2)^{m-j}m!x^{2j}/j!$.

   The initial condition yields $c = \sqrt{e}[1 - \sum_{j=0}^{m}(-2)^{m-j}m!/j!]$.

5. A **more direct method** to find a **particular solution** is to use **undetermined coefficients**.

   We try a mth degree polynomial in x^2 as a particular solution: $u_p = \sum_{j=0}^{m} a_j x^{2j}$, insert in the ODE and find equations for the coefficients a_j:

   $u_p' + xu_p = \sum_{j=0}^{m} 2ja_j x^{2j-1} + \sum_{j=0}^{m} a_j x^{2j+1} = \sum_{j=0}^{m}[(2j+2)a_{j+1} + a_j]x^{2j+1} = x^{2m+1}$ with $a_{m+1} = 0$.

   Then $a_m = 1$ and all other coefficients satisfy: $a_j = -2(j+1)a_{j+1}$, $j = 0, 1, \ldots, m-1$. This gives the particular solution $u_p = \sum_{j=0}^{m}(-2)^{m-j}m!x^{2j}/j!$ and adding the general solution of the homogeneous ODE yields the same general solution as before: $u(x) = ce^{-x^2/2} + \sum_{j=0}^{m}(-2)^{m-j}m!x^{2j}/j!$.

III. FIRST-ORDER NONLINEAR ODE

Most nonlinear equations cannot be solved exactly. However there are classes of equations that can be solved exactly and it is important to know them. The usual procedure is to make a substitution which converts these equations into linear or exactly solvable ones.

A. Bernoulli equations

$$u' = a(x)u + b(x)u^P. \tag{12}$$

For $P = 0, 1$ these equations are **linear** and we already have seen how to solve them. For any other number P these **nonlinear** equations can be converted in linear ones dividing them by u^P and observing that the resulting equation is a linear equation for $y = u^{1-P}$: since $y' = (1-P)u^{-P}u'$, we get

$$y' = (1-P)a(x)y + (1-P)b(x). \tag{13}$$

This equation is linear and inhomogeneous.

Example.

The ODE $u' = x/(x^2u^2 + u^5)$ is not a Bernoulli equation in $u(x)$. However we can rewrite it as $dx/du = u^2x + u^5/x$ which is of Bernoulli type with $P = -1$ for $x(u)$.

The solution is $x(u) = \pm\sqrt{ce^{2u^3/3} - u^3 - \frac{3}{2}}$.

B. Riccati equations

$$u' = a(x)u^2 + b(x)u + c(x). \tag{14}$$

When $a = 0$ this ODE is linear and when $c = 0$ is a Bernoulli equation.

In all other cases the solution can be found analytically if we are able to spot a particular solution of the Riccati equation, no matter how simple that may be. Let $u = u_1(x)$ be a particular solution of (14). The transformation $u = u_1(x) + y(x)$ eliminates $c(x)$ thereby yielding a Bernoulli equation for $y(x)$ that can be solved exactly.

The Bernoulli equation for $y(x)$ is

$$y' = [b(x) + 2a(x)u_1(x)]y + a(x)y^2. \tag{15}$$

Example.

$u' = u^2 - xu + 1$ has the particular solution $u_1 = x$.

The general solution is found by substituting $u = x + y(x)$ in the Riccati equation thereby getting $y' = xy + y^2$.

Division by y^2 gives $-(1/y)' = 1 + x/y$, i.e., $z' + xz = -1$ for $z = 1/y$.

Using the integrating factor $e^{x^2/2}$, we find $(e^{x^2/2}z)' = -e^{x^2/2}$, from which $1/y = z = ce^{-x^2/2} - \int_0^x e^{-(x^2-t^2)/2}dt$.

The solution of the original Riccati equation is

$$u(x) = x + \frac{e^{x^2/2}}{c - \int_0^x e^{t^2/2}dt}.$$

Note: In most cases, a particular solution of a Riccati ODE is not known.
In fact, the substitution

$$u(x) = -\frac{w'(x)}{a(x)w(x)} \tag{16}$$

transforms the Riccati ODE (14) into the second-order linear homogeneous ODE:

$$w'' - \left[\frac{a'(x)}{a(x)} + b(x)\right] w + a(x)c(x)w = 0. \tag{17}$$

This transformation also goes in reverse. There is a Riccati ODE for every second-order homogeneous linear ODE.
Since there is no closed form solution for all second-order linear ODEs, many Riccati ODEs do not have a solution in closed form.

C. Exact equations

These ODEs can be written in the form

$$M(x,u) + N(x,u)u' = \frac{d}{dx} f(x, u(x)) = 0 \tag{18}$$

and the **solution is** $f(x, u(x)) = c$.
A necessary and sufficient **condition** for exactness is that

$$\frac{\partial M}{\partial u} = \frac{\partial N}{\partial x}. \tag{19}$$

Examples.

1. ***Separable ODEs*** are exact because they have the form $M(x) + N(u)u' = 0$. Thus $\partial M/\partial u = \partial N/\partial x = 0$.

2. The ODE $u' = (x^2 - u)/(u^2 + x)$ is ***exact***:
$(u - x^2) + (u^2 + x)u' = 0$ gives $M = u - x^2$, $N = u^2 + x$ so that $\partial M/\partial u = \partial N/\partial x = 1$.
To solve it, we use for example $\partial f/\partial u = N = u^2 + x$.
Then $f(x,u) = \frac{1}{3}u^3 + xu + g(x)$, where $g(x)$ is the "**constant of integration**".
Insertion in $\partial f/\partial x = u + g'(x) = u - x^2$ (which is M) gives $g' = -x^2$ so that $f(x,u) = xu + (u^3 - x^3)/3$ and the **general solution** is $u^3 + 3xu - x^3 = c_1$.

3. ***Integrating factor***. Sometimes multiplication by a factor renders exact a given ODE.

   $(1 + xu + u^2) + (1 + xu + x^2)u' = 0$ **is not exact** because $\partial M/\partial u \neq \partial N/\partial x$.

   However **it becomes exact** when we multiply it by $I = e^{xu}$.

   With this integrating factor, the ODE can be written as $[(x + u)e^{xu}]' = 0$ whose solution is $(x + u)e^{xu} = c$

D. Substitutions

Sometimes a **substitution** converts a nonlinear ODE to one that is directly solvable. Linear substitutions are the easiest to spot.

Some examples:

1. $y = x + u$ converts $u' = \cos(x + u)$ in the **separable ODE**: $y' = 1 + \cos y$.

2. The **linear transformation** $x = av + bw + c$, $u = dv + ew + f$, with a suitable choice of the **coefficients** a, b, c, d, e, f, converts $u' = (Ax + Bu + C)/(Dx + Eu + F)$ into a **separable ODE** for $w(v)$.

3. For an ODE $u' = F(u/x)$, the substitution $y = u/x$ gives a **separable ODE** for $y(x)$: $y' = [F(y) - y]/x$.

4. $u' = u/x + 1/(u + x)$ becomes a **Bernoulli ODE** after the change $y = x + u$: $y' = y/x + 1/y$.

   Setting $z = y^2$, we obtain the linear ODE $z' = 2z/x + 2$.

   A **particular solution** of the form $z = kx$ gives $k = 2k + 2$ or $k = -2$.

   The general solution is then $-2x$ plus a general solution of the homogeneous ODE which is cx^2: $z = cx^2 - 2x$. (Equivalently, divide by x^2 to obtain $(z/x^2)' = 2/x^2$). The solution of the original ODE is $u(x) = -x \pm (cx^2 - 2x)^{1/2}$.

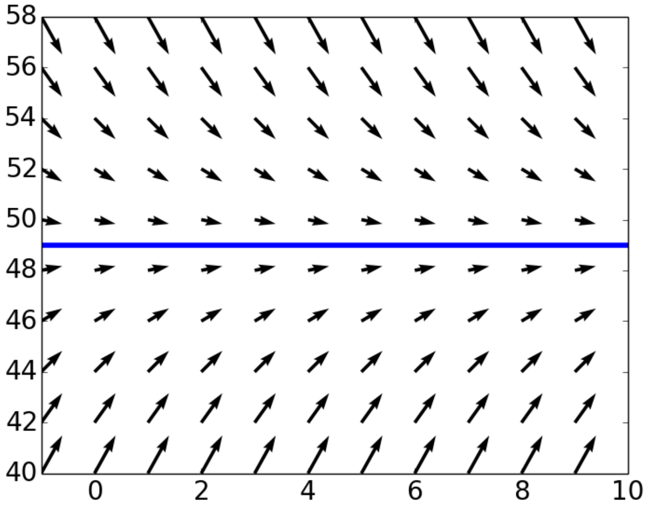

FIG. 1: Direction field in the (x, y) plane for the ODE $y' = 9.8 - 0.2y$ showing the equilibrium solution $y^* = 9.8/0.2 = 49$.

IV. DIRECTION FIELDS, EXISTENCE AND UNIQUENESS OF SOLUTIONS OF THE IVP

Consider the ODE

$$\frac{dy}{dx} = f(x, y). \tag{20}$$

We want to **visualize the solutions** as **trajectories** in the plane (x, y). For that we can cover the plane with a grid and **plot tangent lines of slope** $f(x, y)$ at each grid point (x, y).

This visualization uses the ***computer*** as a tool and produces the ***direction field*** (also called **slope** or **tangent field**) of Fig. 1 for the ODE $y' = 9.8 - 0.2y$.

Note that the grid consists of **points** $x_j = 0.5j$, $j = 0, 1, \dots, 20$ and $y_k = k$, $k = 39, \dots, 60$ that cover the rectangle with corners $(0, 39)$, $(10, 39)$, $(0, 60)$, $(10, 60)$.

The **equilibrium solution** coincides with the ***isocline*** of zero slope $9.8 - 0.2y = 0$, i.e., with the line $y = 49$.

The trajectories $y = y(x)$ are curves whose tangent at a point (x, y) are $f(x, y)$. Visualizing the tangent field, we can **plot qualitatively** the trajectories.

In this example, the trajectories are curves that approach the equilibrium solution as $x \to \infty$. This of course agrees with the fact that the solution of the ODE parametrized by the arbitrary integration constant c is

$$y(x) = 49 + c\, e^{-x/5}.$$

Equivalently, the solution of the *initial value problem* (IVP) $y = y_0$ at $x = x_0$ is

$$y(x) = 49 + (y_0 - 49)\, e^{-(x-x_0)/5}.$$

From the slope field, we expect that the solution that starts at a given point (x_0, y_0) is not crossed by any other trajectory. In fact, at point where two different trajectories cross there would be two different values of $f(x, y)$ which cannot occur for a single-valued function. This condition is made more precise by the following:

Theorem of existence and uniqueness:

Provided f and $\partial f/\partial y$ are continuous in a rectangle R and $(x_0, y_0) \in R$, $y' = f(x, y)$ has a unique solution for $|x - x_0| < \delta$ (for some $\delta > 0$ that leaves x in the rectangle R) that satisfies the initial condition $y(x_0) = y_0$.

The **proof** can be found in most books. See for instance page 112 of Boyce and Di Prima [1]. Typically the IVP is transformed in an integral equation that is solved by iteration justifying the steps. It is possible to prove that the solutions depend continuously on the initial data (x_0, y_0) although that is outside the scope of this course. See chapter 2 of [4].

What happens if the premises of the theorem are not fulfilled?

Consider the IVP

$$\frac{dy}{dx} = 2\sqrt{y}, \quad y(0) = 0.$$

Clearly the premises of the theorem are not satisfied at $(0, 0)$ because $\sqrt{y}$ does not have a continuous derivative there.

Separation of variables gives the solution of the IVP $y(x) = x^2$, but $y(x) = 0$ and $y(x) = (x - \xi)\theta(x - \xi)$ for any $\xi > 0$ are also solutions. [$\theta(x) = 1$ for $x > 0$ and $\theta(x) = 0$

for $x < 0$ is the **Heaviside** unit step function.]

It turns out that having a continuous $f(x, y)$ at (x_0, y_0) and in a rectangle about it guarantees the existence of a solution to the IVP but, as we have seen with the previous example, we need continuity of $\partial f/\partial y$ to have a unique solution of the IVP.
A slightly more precise and general condition can be found in the version of the existence and uniqueness theorem of [4].
We now calculate the **slope field of a more difficult example**:

$$\frac{dy}{dx} = xy(y-2). \tag{21}$$

The computer construction of the slope field proceeds by selecting a large enough rectangle, setting a grid and depicting the slopes at the grid points.
The *human* construction has to be somewhat more intelligent. We want to separate the plane in sectors where we know the sign of y' and of y''. In particular, we want to find sectors where:

a. $y' < 0$, $y'' < 0$: trajectories are decreasing and concave;

b. $y' < 0$, $y'' > 0$: trajectories are decreasing and convex;

c. $y' > 0$, $y'' < 0$: trajectories are increasing and concave;

d. $y' > 0$, $y'' > 0$: trajectories are increasing and convex.

Sectors of equal sign of y' are separated by ***isoclines*** of slope zero (**horizontal tangent)** or infinite (**vertical tangent**).
Sectors of equal sign of y'' are separated by **lines of inflection**, points where $y'' = 0$.
We have to use the chain rule to calculate y'':

$$\frac{d^2y}{dx^2} = \frac{d}{dx}f(x, y(x)) = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\frac{dy}{dx} = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}f.$$

Now we follow the **protocol**:

i) Draw the zero-slope isoclines (also called **nullclines**): $f(x, y) = 0$;

ii) Draw the infinite-slope isoclines: $1/f(x, y) = 0$;

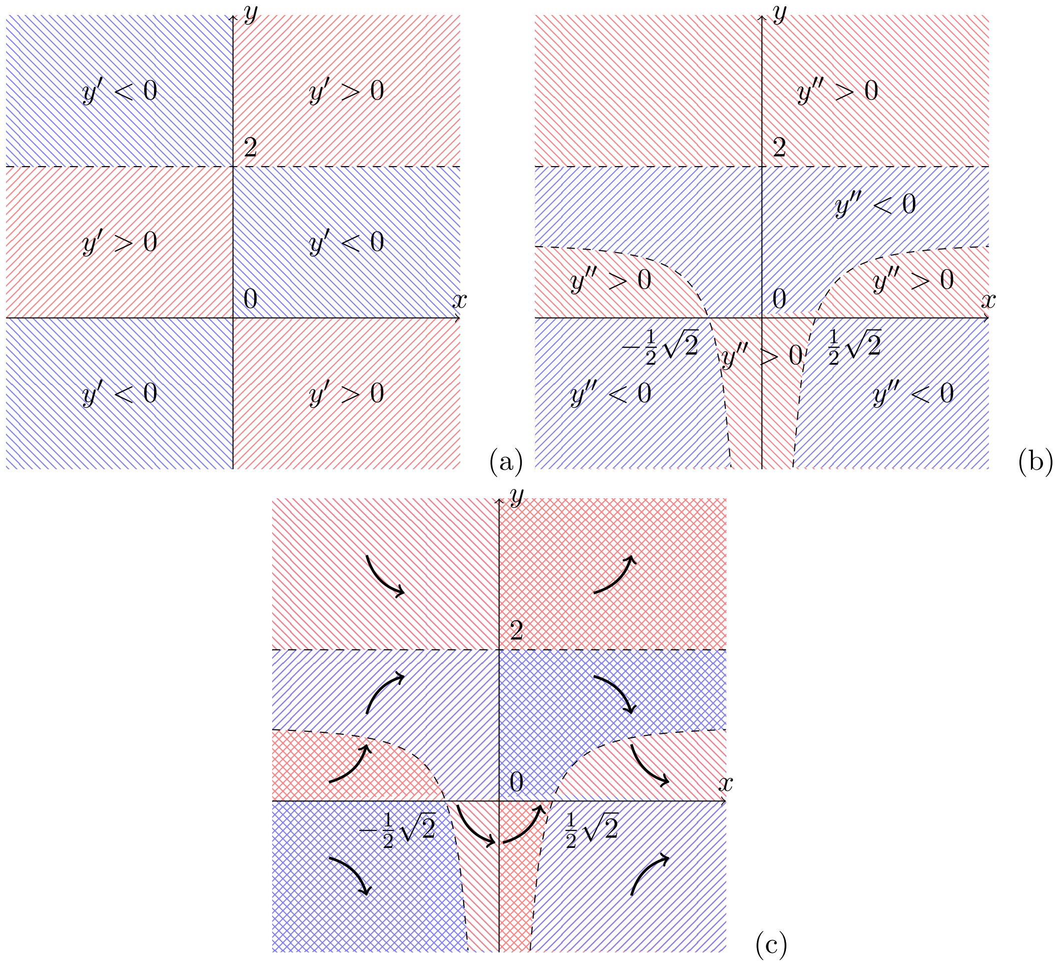

FIG. 2: Sectors of (a) decreasing and increasing $y(x)$, (b) concave and convex $y(x)$, (c) combined information about signs of y' and y''. The ODE is $y' = xy(y-2)$.

iii) Draw the inflection point curves: $y'' = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} f = 0$;

iv) Separate the plane in the sectors a, b, c, d as written above.

For the example (21), we find:

i) Nullclines: $x = 0$, $y = 0$, and $y = 2$. The nullclines $y = 0$ and $y = 2$ are also lines of equilibrium points.

ii) No lines of vertical slope.

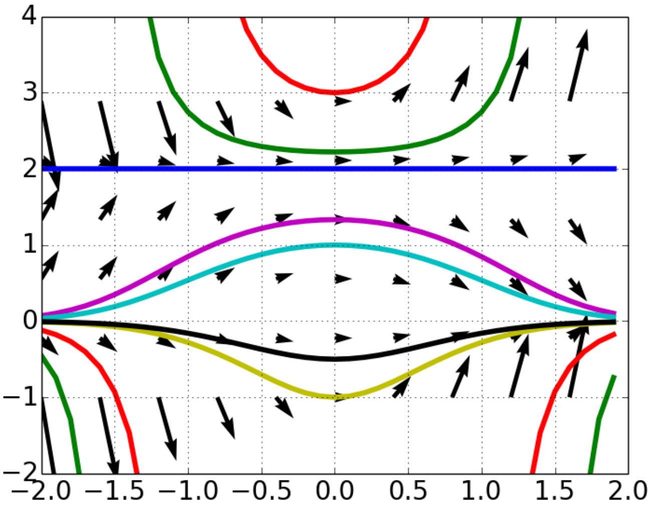

FIG. 3: Slope field and trajectories in the (x, y) plane for the ODE $y' = xy(y-2)$.

iii) $y'' = y(y-2)(1+2x^2y-2x^2) = 0$. Thus the lines of inflection points are $y=0$, $y=2$ and $y = 1 - \frac{1}{2x^2}$.

iv) The separation of the plane in sectors is as indicated in Figure 2.

Figure 3 shows the slope field and several representative trajectories.

Separating variables, we find the general solution:

$$\int x\,dx = \int \frac{dy}{y(y-2)} = \frac{1}{2}\int \left[\frac{1}{y-2} - \frac{1}{y}\right] dy = \frac{1}{2}\ln\left|\frac{y-2}{y}\right| = \frac{1}{2}\ln\left|1-\frac{2}{y}\right|.$$

This gives $|1-2/y| = c\,e^{x^2}$, i.e. $1 - 2/y = \pm c\,e^{x^2}$ $(c>0)$ or

$$y = \frac{2}{1-c\,e^{x^2}}, \qquad c \text{ es un número real arbitraro.} \tag{22}$$

The trajectories of the figure 3 are obtained by taking the following values of the constant c in equation number (22):

(red) $c = \frac{1}{3}$, (green) $c = 0.1$, (magenta) $c = -\frac{1}{2}$, (cyan) $c = -1$, (blue) $c = 0$, (yellow) $c = 3$.

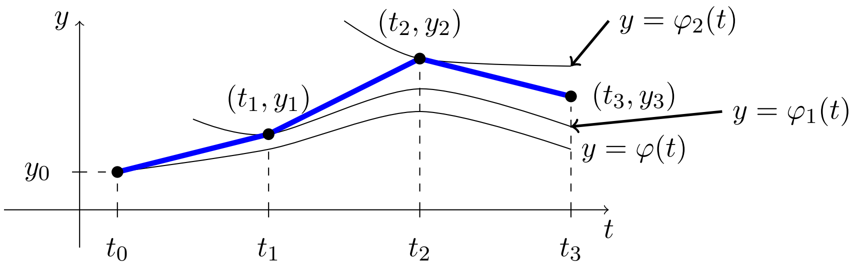

FIG. 4: The Euler method approximates a given trajectory $y = \varphi(t)$ by segments that move it to nearby trajectories $\varphi_1(t)$, $\varphi_2(t)$, ... in the (t, y) plane.

V. NUMERICAL METHODS: EULER METHOD

The oldest numerical method to solve the general first-order IVP (20), $y' = f(t, y)$ with $y(t_0) = y_0$, uses the slope field [1].

In fact, at the point t_0, **the tangent line to the trajectory passing through this point is**

$$y = y_0 + f(t_0, y_0)(t - t_0).$$

We can approximate the trajectory by this tangent for small enough $|t - t_0|$, up to a point t_1 so that $y(t_1) \approx y_1 = y_0 + f(t_0, y_0)(t_1 - t_0)$. After the segment $t_1 - t_0$, we can approximate $y(t) \approx y_1 + f(t_1, y_1)(t - t_1)$ up to a point t_2 so that $y(t_2) \approx y_2 = y_1 + f(t_1, y_1)(t_2 - t_1)$, and so on.

The type of approximation we are doing is depicted in Fig. 4. If we divide a time interval $[0, T]$ in N equal pieces of width h, we obtain the usual form of the **Euler scheme** that starts at the initial condition $y(t_0) = y_0$:

$$\begin{aligned} & y_{j+1} = y_j + f(t_j, y_j)\, h, \\ & t_0 < t_1 = t_0 + h < \ldots < t_j = t_0 + jh < \ldots < t_N = t_0 + Nh = T. \end{aligned} \tag{23}$$

The Euler method approximates better the solution as the segments $|t_{j+1} - t_j|$ become smaller.

From Fig. 4 we also conclude that when the trajectories are converging the errors are small whereas they increase when the trajectories are diverging.

A **more quantitative assessment of the Euler method** [1], can be obtained as follows. First we rewrite the IVP as the equivalent ***integral*** **equation**:

$$y(t) = y_0 + \int_0^t f(y(s), s)\, ds. \tag{24}$$

This equation immediately produces for $0 \leq \tau < t \leq T$:

$$y(t) = y(\tau) + \int_\tau^t f(y(s), s)\, ds. \tag{25}$$

To approximately solve this equation on an time interval $[0, T]$, we divide this interval in N pieces: $0 \leq t_1 = h, \ldots, t_N = hN$ and approximate the integral by some **quadrature scheme**. One-step methods set $t = t_{j+1}$ and $\tau = t_j$.

The simplest such methods is the **Euler scheme**:

$$y_{j+1} = y_j + hf(t_j, y_j), \tag{26}$$

in which the integral is approximated by h times the integrand evaluated at the earlier time t_j.

We have called y_j the approximation to the solution $y(t)$ at time $t = t_j = jh$.

Note that the Euler method also follows from approximating the derivative $dy(t)/dt \approx (y_{j+1} - y_j)/h$.

For general **one-step schemes**, we can prove that **consistent** one-step methods are also **convergent**.

Example 1.

Consider the IVP

$$\begin{cases} \frac{dy(t)}{dt} = ry(t), \\ y(0) = u_0 > 0. \end{cases} \tag{27}$$

to be solved on $[0, 1]$.

The **Euler method** gives

$$y_{j+1} = y_j + hry_j = (1 + hr)y_j, \quad y_0 = y_0, \quad h = \frac{1}{N}, \quad j = 0, 1 \ldots, N. \tag{28}$$

The solution of the scheme is

$$y_j = (1 + hr)^j y_0. \tag{29}$$

Note that the **exact solution** is $y(t) = e^{rt}y_0$ so that $y(t_j + h) = e^{rh}y(t_j)$.

Equation (28) shows that the Euler method approximates e^{rh} by the straight line $1 + rh$.

Errors.

Suppose that we have approximated the IVP by a **one-step numerical scheme** $y_{j+1} = y_j + \Phi(t_j, y_j, y_{j+1}, h)\, h$.

We have $\Phi(t_j, y_j, y_{j+1}, h) = f(t_j, y_j)$ for the Euler scheme. The latter is an ***explicit* scheme** because y_{j+1} is given as an explicit function of y_j.

Let us define the local ***truncation error*** or **discretization error** of the one-step numerical scheme as:

$$\tau_{j+1} = \left| \frac{y(t_{j+1}) - y(t_j)}{h} - \Phi(t_j, y(t_j), y(t_{j+1}), h) \right|, \quad j = 0, 1, \ldots, N-1, \tag{30}$$

where we substitute the exact solution $y(t_j)$ instead of y_j.

If the τ_j vanish as $h \to 0$, we say that **the difference equations are *consistent* with the differential equation**.

For the Euler method (23), the Taylor theorem and the chain rule give

$$\begin{aligned} \frac{y(t_{j+1}) - y(t_j)}{h} &= \frac{y(t_j + h) - y(t_j)}{h} = \frac{dy(t_j)}{dt} + \frac{h}{2}\frac{d^2y(\xi)}{dt^2} \\ &= f(y(t_j), t_j) + \frac{h}{2}\left[\frac{\partial f}{\partial t}(\xi, y(\xi)) + \frac{\partial f(\xi, y(\xi))}{\partial y} f(\xi, y(\xi))\right], \end{aligned} \tag{31}$$

where $t_j \le \xi \le t_{j+1}$, provided continuous partial derivatives of f exist.

In this case, **the maximum truncation error** is bounded by

$$\tau = \max_j |\tau_j| \le \frac{h}{2} M_2, \quad M_2 = \sup_{0 \le t \le T} \left| \frac{d^2 y(t)}{dt^2} \right|. \tag{32}$$

The result (32) indicates that **the Euler method** has local **truncation error** of **order** h.

Definition. We say that $f(h) = O(g(h))$ as $h \to 0$ if there exist two positive constants c and ϵ such that $|f(h)| < c|g(h)|$ for all $|h| < \epsilon$.

It is possible to show that **the global error** $\sup_j |y(t_j) - y_j|$ is also of order h and this shows that **the Euler method** is an $O(h)$ or ***first-order* method**.

Code.

We will implement Euler's method with *Matlab* to integrate the following problem:

$$\begin{cases} y' + y = 0 \\ y(0) = 1 \end{cases},$$

on the domain $0 \leq x \leq 5$. We will use several discretization steps ($h = 0.5$, 0.25, 0.1) and compare the approximate solution with the exact solution $y(x) = e^{-x}$. A *Matlab* code for the Euler method can be the following:

```
clear all  % To clear all the previous data .
h=0.5;  % The step of the discretization.
a=0;b=5;  % Domain.
N=round((b-a)/h)  % N+1 is the number of total nodes.
x(1)=0 ; y(1)=1;  % Initial condition.
exacsol(1)=y0;  % First value of the exact solution.

for i= 1:N  % Onset of the iterative process.
    y(i+1)=y(i)+h*(-y(i));%EXPLICIT EULER ESCHEME
    x(i+1)=x(i)+h; %Next node.
    exacsol(i+1)=exp(-x(i+1));
end

plot(x,y,'r*',x,exacsol)
xlabel('x') ; ylabel('y') ; legend('Aprox. by explicit Euler','Exact solution')
```

Let us explore the relationship between the order of the Euler method and the results obtained using it.

Table I compares the results obtained by the Euler method with steps $h = 0.5$, 0.25, 0.1. The *total error* is the maximum error made by the method as compared to the exact solution. For the Euler method, these errors decrease as h : if we take $h_1 = 0.5$, the error is $error_1 = 0.1179$. Reducing now the step by factors 2 and 5 (i.e., $h_2 = h_1/2 = 0.25$ and $h_3 = h_1/5 = 0.1$), reduces the error by factors 2 and 5, respectively ($error_1/2 \approx error_2 = 0.0515$, $error_1/5 \approx error_3 = 0.0192$).

h	N (number of nodes)	Total error of the Euler method
0.5	11	0.1179
0.25	21	0.0515
0.1	51	0.0192

TABLE I: Total errors of the Euler method for steps $h = 0.5$, 0.25 and 0.1.

Backward Euler method.

We could have approximated the integral in (25) for $t = t_{j+1}$ and $\tau = t_j$ by $hf(t_{j+1}, y_{j+1})$, thereby getting

$$y_{j+1} = y_j + hf(t_{j+1}, y_{j+1}), \quad j = 0, 1, \ldots, N-1. \tag{33}$$

This scheme is ***implicit*** because the right-hand side is a function of y_{j+1} (unless f is independent of y) and we have to solve (33) for y_{j+1} at each step. (33) is called the backward Euler method and it could have been obtained from the ODE $y' = f(t, y)$ had the time derivative dy/dt at time t_{j+1} been replaced by the backward finite difference $(y_{j+1} - y_j)/h$. The backward Euler method is also an $O(h)$ method.

Example 2.

Solve the first-order IVP by the backward Euler method.

We get

$$y_{j+1} = y_j + hry_{j+1}, \quad y_0, \quad h = \frac{1}{N}, \quad j = 0, 1 \ldots, N-1, \tag{34}$$

insead of (28). Solving this equation for y_{j+1}, we find

$$y_{j+1} = \frac{y_j}{1 - hr} = \frac{y_0}{(1 - hr)^j}, \quad j = 0, 1 \ldots, N-1. \tag{35}$$

This equation makes sense for h small enough so that $hr < 1$.

Example 3: failure of the Euler method.

Solving the IVP $y' = 2\sqrt{y}$, $y(0) = 0$, by the Euler method, we only find the solution $y(0) = 0$. The reason is that the Euler method is explicit and if we start with $y_0 = 0$, we always obtain $y_j = 0$.

On the other hand, the backward Euler method gives $y_{j+1} = y_j + 2hy_{j+1}^{1/2}$.

For $y_0 = 0$, we find $y_1 = 2hy_1^{1/2}$, which has the solutions $y_1 = 0$ and $y_1^{1/2} = 2h$ or $y_1 = (2h)^2$. If we have reached $y_1 = (2h)^2$, the next iterates try to reach the solution of the ODE, $y(t) = t^2$. Any small error that makes $y_0 > 0$ will give iterates trying to reach $t(t) = t^2$.

VI. HEUN'S METHOD

The Euler method is simple but low order. To improve it, we need to approximate better the integral in (24). A possibility is to use the **trapezoidal rule**

$$\int_{t_j}^{t_{j+1}} f(s, y(s))\, ds \approx \frac{h}{2}\left[f(t_{j+1}, y(t_{j+1})) + f(t_j, y(t_j))\right]. \tag{36}$$

We obtain the following **implicit method**

$$y_{j+1} = y_j + \frac{h}{2}[f(t_{j+1}, y_{j+1}) + f(t_j, y(t_j))], \quad j = 0, 1, \ldots, N-1, \tag{37}$$

which is sometimes called a **Crank-Nicholson scheme**.
To make this scheme explicit, we can calculate $y(t_{j+1})$ in (36) using the Euler method. The result is called ***Heun's method***:

$$y_{j+1} = y_j + \frac{h}{2}[f(t_{j+1}, y_j + hf(y_j, t_j)) + f(t_j, y_j)], \quad j = 0, 1, \ldots, N-1. \tag{38}$$

Heun's method is an example of a ***predictor-corrector*** **scheme** in which we use Euler's method to make a prediction of $y(t_{j+1})$ and correct it by inserting this prediction in (36). (38) may be equivalently written as

$$\begin{aligned} p_{j+1} &= y_j + hf(t_j, y_j), \quad t_{j+1} = t_j + h, \\ y_{j+1} &= y_j + \frac{h}{2}[f(t_{j+1}, p_{j+1}) + f(t_j, y_j)], \quad j = 0, 1, \ldots, N-1. \end{aligned} \tag{39}$$

Using the Taylor theorem, we can prove that Heun's method has a **truncation error** $\tau = O(h^2)$. The **global error** is also of **order** h^2.
Accordingly, if we halve the step in Heun's method, the global error becomes one quarter.

Example 4.
Heun's method applied to the IVP (27) of Example 1 yields

$$\begin{aligned} p_{j+1} &= y_j + hry_j, \\ y_{j+1} &= y_j + \frac{hr}{2}(p_{j+1} + y_j), \quad j = 0, 1 \ldots, N-1. \end{aligned} \tag{40}$$

In this case, $p_{j+1} = (1 + hr)y_j$ and (40) produces

$$y_{j+1} = \left(1 + hr + \frac{h^2 r^2}{2}\right) y_j = \left(1 + hr + \frac{h^2 r^2}{2}\right)^j y_0, \quad j = 0, 1 \ldots, N-1. \tag{41}$$

The **exact solution** is $y(t_j + h) = e^{rh} y(t_j) = e^{r(t_j+h)} y_0$ which compared to (41) shows that the Heun method produces one more term in the approximation of the exponential e^{rh} than the Euler method that gives $y_{j+1} = (1 + rh)y_j = (1 + rh)^j y_0$.

VII. RUNGE-KUTTA METHODS

A different family of integration methods for the usual IVP starts by Taylor expanding a finite difference:

$$\begin{aligned} y(t+h) - y(t) &= h\frac{dy}{dt}(t) + \frac{h^2}{2}\frac{d^2 y}{dt^2}(t) + O(h^3) \\ &= hf(t,y) + \frac{h^2}{2}\left[\frac{\partial f}{\partial t}(t, y(t)) + \frac{\partial f}{\partial y}(t, y(t))\, f(t, y(t))\right] + O(h^3), \end{aligned} \tag{42}$$

where we have substituted $\frac{\partial f}{\partial t}(t, y(t)) + \frac{\partial f}{\partial y}(t, y(t))\, f(t, y(t))$ instead of $\frac{d^2 y}{dt^2}(t)$.
Ignoring the error term, we obtain a **numerical scheme of order 2**.
However having to carry out the partial derivatives is computationally costly. Thus the Runge-Kutta (RK) idea is to replace the right hand side of (42) by a linear combination of two functions that gives a related scheme of the same order:

$$y(t+h) = y(t) + Ahf_0 + Bhf_1, \quad f_0 = f(t,y), \quad f_1 = f(t + Ph, y + Qhf_0).$$

We now use the Taylor formula for functions of two variables to expand f_1:

$$y(t+h) = y(t) + (A+B)hf(t,y) + BPh^2\frac{\partial f}{\partial t}(t,y) + BQh^2\frac{\partial f}{\partial y}(t,y)\, f(t,y) + O(h^3).$$

Equating this to (42), we obtain

$$A + B = 1, \quad BP = \frac{1}{2}, \quad BQ = \frac{1}{2}. \tag{43}$$

This is a system of 3 equations for 4 unknowns, A, B, P and Q. Therefore we can give values to one of the unknowns and obtain different **second order RK schemes (RK2)**.
Common RK2 schemes are

1. $A = \frac{1}{2}$ and therefore $B = \frac{1}{2}$, $P = 1$, $Q = 1$. We recover the **Heun scheme**:

$$y(t+h) = y(t) + \frac{h}{2}\left[f(t,y) + f(t+h, y+hf(t,y))\right].$$

2. $A = 0$ gives $B = 1$, $P = Q = \frac{1}{2}$ and the ***mid-point*** RK2, also called **modified Euler method** or **Cauchy method**, is obtained:

$$y(t+h) = y(t) + h\,f\left(t + \frac{h}{2}, y + \frac{h}{2}f(t,y)\right).$$

3. $A = \frac{1}{4}$ gives $B = \frac{3}{4}$, $P = Q = \frac{2}{3}$ and we find

$$y(t+h) = y(t) + \frac{h}{4}f(t,y) + \frac{3h}{4}\,f\left(t + \frac{2h}{3}, y + \frac{2h}{3}f(t,y)\right).$$

It is possible to prove that this choice gives a truncation error $O(h^4)$.

Similar ideas are used to generate more precise RK3 and RK4 schemes commonly used in numerical codes.

The Matlab routine ***ode45*** is a variable-step RK4 scheme whose step size is adjusted using a RK5 scheme to estimate the error made after each step.

[1] W.E. Boyce and R.C. Di Prima, Elementary differential equations and boundary value problems. 9th ed. John Wiley & Sons, N.Y. 2009.

[2] G.F. Carrier and C.E. Pearson, Ordinary differential equations. SIAM Classics in Applied Mathematics **6**. SIAM, PA 1991.

[3] G. F. Simmons, Differential equations with applications and historical notes. 2nd ed. MacGraw Hill, N.Y. 1991. Translation into Spanish appeared in 1993.

[4] R. M. M. Mattheij and J. Molenaar, Ordinary differential equations in theory and practice. Classics in Applied Mathematics **43**. SIAM, PA 2002. This book has more precise mathematical content than this lecture but its level is higher and it is harder to read.

[5] C. M. Bender and S. A. Orszag, Advanced Mathematical Methods for Scientists and Engineers. McGraw Hill, N.Y. 1978. This is also an advanced book but Chapter 1 contains a useful compendium of methods to find exact solutions of ODEs.