# Ordinary Differential Equations

Adrien Deloro

AIMS Sénégal, 27th November – 15th December 2017

# Contents

# Introduction

### Prerequisites

In order to follow this class, one needs to know:

- calculus: derivatives and integrals;
- analysis: the same, with the epsilons;
- linear algebra: matrices, but also vector spaces and linear transformations.

### Recommended Reading

I could locate two relevant sources. One is a classical book; l'autre est un polycopié de l'université d'Orsay.


[1] Robert Devaney, Morris Hirsch, and Stephen Smale. *Differential equations, dynamical systems and an introduction to chaos*. Elsevier/Academic Press, Amsterdam. Any of the three editions.

[2] Dominique Hulin. *Équations differentielles ordinaires, études qualitatives*. Téléchargeable depuis https://www.math.u-psud.fr/~hulin/poly-cours-EDO.pdf


## Differential equations everywhere

What is a differential equation? It is an equation:

- where the unknown variable is a function;
- involving the function and its derivatives.

And before we say anything else, the reader must be warned that *to solve is not the same thing as to study*: as we shall see, most differential equations cannot be solved explicitly by an exact formula. But they can certainly be studied.

What are differential equations for? They can be used to model mathematically every natural phenomenon where time plays a role — more precisely, continuous (real-valued) time. This covers quite a lot as we shall see. On the other hand they cannot be used to model the discrete (integer-valued) time computers follow. But computers can help us study differential equations; the interplay between problems in continuous and discrete time is extremely subtle, and the subject of *numerical analysis*. We shall not touch this. Our purpose in this first lecture is to demonstrate that differential equations are (almost) everywhere.

## Gravity

Drop a piece of chalk: it will fall.

According to Newton's law of gravitation, and as a first approximation, the stick of chalk is subject to a force $\overrightarrow{F_{\text{grav}}} = m\vec{g}$, where:

- $m$ is the mass;
- $\vec{g}$ is the gravitation vector near earth surface.

We neglect other forces such as air resistance.

On the other hand, by Newton's second law, $\sum \vec{F} = m\vec{a}$, where the sum ranges over all forces involved, and $\vec{a}$ is the acceleration vector.

The latter is the second derivative of the position vector. Let us be more specific. Say that at instant $t$, the piece of chalk has position:

$$\begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix};$$

the three coordinates are functions of the time $t$. However motion is vertical, so $x$ and $y$ are actually constants in this example. Now the acceleration is:

$$\vec{a}(t) = \begin{pmatrix} 0 \\ 0 \\ z''(t) \end{pmatrix}$$

Since vector $\vec{g}$ is a constant, vertical vector pointing down, it has the simple expression

$$\vec{g} = \begin{pmatrix} 0 \\ 0 \\ -g \end{pmatrix}$$

Put toghether and keeping only the $z$-coordinate, we find:

$$z''(t) = -g$$

Our first differential equation!

Let us introduce some terminology:

- we call $z''(t) = -g$ a <u>scalar</u> equation, because $z : \mathbb{R} \to \mathbb{R}$ is a real-valued function (instead of a vector-valued function);
- it is a <u>linear</u> equation because it can be written:

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

  where $z^{(k)}$ stands for the $k^{\text{th}}$ derivative.

**Remark.** Instead of *linear*, one should call the equation affine for consistency with geometry. In geometry the equation $y + ax = 0$ is linear, whereas equation $y + ax = b$ is merely affine.

This unfortunate misuse of "linear" in the theory of differential equations is the source of many confusions.

- The equation has constant coefficients because the $a_k$'s above are actually constant: $z'' + 0z' + 0z$; *although here $b(t) = g$ is another constant, it is irrelevant*;
- it has order 2 because the highest derivative involved is $z''$.

## Chemistry

A liquid (what the chemists call a solution) contains molecules of types $Y_1, Y_2, Y_3$ with concentrations $c_1(t), c_2(t), c_3(t)$ subject to chemical reactions:

$$Y_3 \xrightarrow{k_3} Y_2 \xrightarrow{k_2} Y_1 \xrightarrow{k_1} Y_2.$$

This means that between instants $t$ and $t+dt$, a certain amount of molecules of type $Y_3$ will transform into molecules of type $Y_2$; this amount is proportional to the concentration and the time elapsed. This can be modelled as:

$$c_3(t + dt) = c_3(t) - k_3 c_3(t) dt,$$

or:

$$\frac{c_3(t + dt) - c_3(t)}{dt} = -k_3 c_3(t).$$

Now if $dt$ is a very brief amount of time, we can call calculus in and write:

$$c_3'(t) = \frac{dc_3(t)}{dt} = -k_3 c_3(t).$$

Likewise, as far as molecules of type $Y_2$ are concerned:

- some molecules of type $Y_3$ turn into molecules of type $Y_2$, which results in an increase of $+k_3 c_3(t)$;
- but some molecules of type $Y_2$ turn into molecules of type $Y_1$, resulting in a decrease $-k_2 c_2(t)$;
- then again some molecules of type $Y_1$ turn into type $Y_2$, which is a gain of $+k_1 c_1(t)$.

Consequently, $c_2'(t) = k_1 c_1(t) - k_2 c_2(t) + k_3 c_3(t)$.

Finally, $c_1'(t) = -k_1 c_1(t) + k_2 c_2(t)$. This gives us the following system of coupled scalar equations:

$$\begin{cases} c_1'(t) & = & -k_1 c_1(t) + k_2 c_2(t) \\ c_2'(t) & = & k_1 c_1(t) - k_2 c_2(t) + k_3 c_3(t) \\ c_3'(t) & = & -k_3 c_3(t) \end{cases}$$

- **The matrix trick**

The previous is a bit tedious to write. Just like linear algebra helps solve linear systems, let us introduce the concentration vector $C(t) = \begin{pmatrix} c_1(t) \\ c_2(t) \\ c_3(t) \end{pmatrix}$. It is then subject to the equation:

$$C'(t) = \begin{pmatrix} -c_1 & k_2 & 0 \\ k_1 & -k_2 & k_3 \\ 0 & 0 & -k_3 \end{pmatrix} \cdot C(t)$$

So our system of chemical equations actually reduces to one vector linear equation of order 1.

**Remark.** More generally, a linear, vector differential equation of order 1 takes the form

$$C'(t) = A(t) \cdot C(t) + B(t);$$

notice that:

- matrix $A(t)$ may vary in general; in our example from chemistry, we therefore have constant coefficients;
- there can be a $+B(t)$ term, with $B(t)$ a variable vector. If there is no such term, i.e. if $B(t) = \vec{0}$ as was the case in the example, physicists call the equation homogeneous.

This is bad terminology as pointed out.

| | what they say | what they should say |
|---|---|---|
| $C'(t) = A(t)C(t) + B(t)$ | linear | affine |
| $C'(t) = A(t)C(t)$ | linear homogeneous | linear |

Unfortunately, tradition is strong.

# Biology

Consider a box containing bacteria with population $p(t)$. Between instants $t$ and $t + dt$,

- a number of new bacteria are born, say $\beta p(t)dt$, where $\beta$ is the birth factor;
- and a number of them die, say $\delta p(t)dt$, where $\delta$ is the death factor.

Put together:

$$p(t + dt) = p(t) + \beta p(t)dt - \delta p(t)dt,$$

so letting $\gamma = \beta - \delta$:

$$p't) = \frac{dp}{dt} = \gamma p(t).$$

This is the Malthus model for population: a scalar, first order, linear equation with constant coefficients.

Of course the Malthus model is not realistic, as finiteness of the box yields constraints. In the middle of the XIXth century Verhulst suggested the following model, which is surprisingly accurate:

$$p'(t) = c \cdot p(t) \cdot \left(1 - \frac{p(t)}{K}\right),$$

which is a scalar, but non-linear equation of order 1. (The equation does contain a term $p^2(t)$, but what matters in computing the order is the highest derivation order, not powers involved.)

**Remark.** The Verhulst equation is one of the few non-linear equations for which solutions can be written down explicitly in terms of the usual functions (it is an excellent exercise to find the formula).

*In general, a non-linear equation cannot be solved by an explicit formula.*

Let us go further: in the 1920's Lotka and Volterra suggested the following prey/predator model:

$$\left\{\begin{array}{lcl} p_1'(t) & = & p_1(t) \cdot (a - bp_2(t)) \\ p_2'(t) & = & p_2(t) \cdot (-c + dp_1(t)) \end{array}\right. ,$$

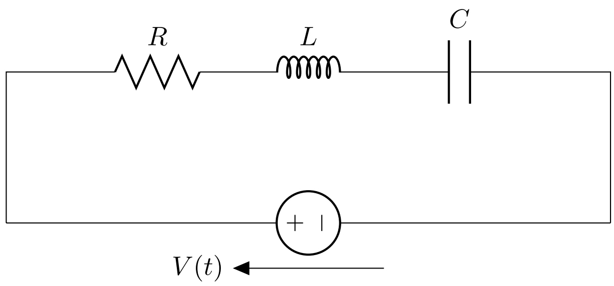
a system of coupled, scalar, non-linear, first order equations (for which there exists no explicit formula). The meaning is simple:

- population $p_1$ consists of rabbits; rabbits tend to reproduce exponentially if it were not for foxes who kill them (the number of casualties being proportional to the number of rabbits and to the number of foxes);
- population $p_2$ consists of foxes; this population is prosperous only if there are rabbits to feed on.

Epidemiology too relies on differential equations.

## Electrical engineering

The RLC circuit stands for Resistor, Iinductor (they use $L$, because $I$ already stands for electric current, in French Intensité), Capacitor. Let us assemble these electrical components in series together with a variable voltage source $V(t)$, as follows:




Let us find the resulting equation. Let $I(t)$ stand for the current and $V_X(t)$ be the entrance voltage at component $X$. Then:

- by Kirchhoff's law: $V_R(t) + V_L(t) + V_C(t) = V(t)$;
- by Ohm's law: $V_R(t)RI(t)$;
- by Faraday's law: $V_L = LI'(t)$;
- by Coulomb's law: $CV_C'(t) = I(t)$.

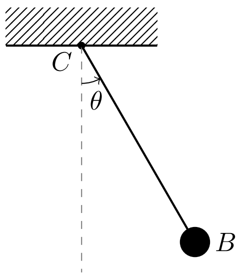
Consequently,

$$RI'(t) + LI''(t) + \frac{1}{C}I(t) = V'(t).$$

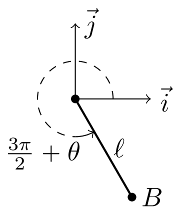
This is a <u>second order, scalar</u> equation; people call it <u>linear</u> (but really, they should say affine) and to make things worse, they say it has <u>constant coefficients</u> even though $V(t)$ depends on time.

## Mechanics

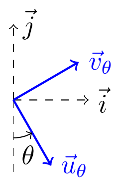
Consider a simple pendulum: a point $B$ (the bob) with mass $m$ hangs from $C$ (the centre) through a massless, inextensible cord of length $\ell$.




The point is located using the angular displacement $\theta(t)$, so that $B$ has polar coordinates $\left(\ell, \frac{3\pi}{2} + \theta(t)\right)$:




In this and similar problems from physics it is often preferable not to use Cartesian $(x, y)$ nor even polar $(\rho, \varphi)$ coordinates, but work in the variable basis $(\vec{u}_\theta, \vec{v}_\theta)$, defined as follows:




- $\vec{u}_\theta$ is the unit vector along $\overrightarrow{CB}$;
- $\vec{v}_\theta$ is obtained from $u_\theta$ by rotating it $+\frac{\pi}{2}$.

Hence:

$$\begin{aligned}\vec{u}_\theta &= (\vec{u}_\theta \cdot \vec{i})\vec{i} + (\vec{u}_\theta \cdot \vec{j})\vec{j} \\ &= \cos\left(\frac{3\pi}{2} + \theta\right)\vec{i} + \sin\left(\frac{3\pi}{2} + \theta\right)\vec{j} \\ &= \sin\theta\vec{i} - \cos\theta\vec{j},\end{aligned}$$

and likewise $\vec{v}_\theta = \cos\theta\vec{i} + \sin\theta\vec{j}$.

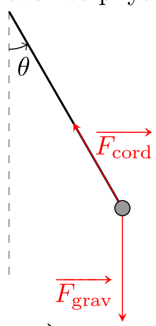
Then notice:

$$\frac{d\vec{u}_\theta}{d\theta} = \vec{v}_\theta \quad \text{and} \quad \frac{d\vec{v}_\theta}{d\theta} = -\vec{u}_\theta,$$

so that using the chain rule:

$$\frac{d\vec{u}_\theta}{dt} = \theta'(t)\vec{v}_\theta \quad \text{and} \quad \frac{d\vec{v}_\theta}{dt} = -\theta'(t)\vec{u}_\theta.$$

Now back to physics. The bob is subject to two forces:




- $\overrightarrow{F_{\text{grav}}} = m\vec{g} = -mg\vec{j} = -mg(-\cos\theta\vec{u}_\theta + \sin\theta\vec{v}_\theta)$;
- $\overrightarrow{F_{\text{cord}}} = -T\vec{u}_\theta$.

Since $\overrightarrow{CB} = \ell\vec{u}_\theta$ and $\ell$ is a constant, one finds the velocity:

$$\overrightarrow{CB}' = \ell\vec{u}_\theta' = \ell\theta'\vec{v}_\theta,$$

and acceleration:

$$\overrightarrow{CB}'' = \ell\theta''\vec{v}_\theta - \ell\theta'^2\vec{u}_\theta.$$

Finally by Newton's second law, $\overrightarrow{F_{\text{grav}}} + \overrightarrow{F_{\text{cord}}} = m\vec{a}$, so taking the scalar product with $\vec{v}_\theta$:

$$-mg\sin\theta = m\ell\theta'',$$

or equivalently

$$\theta'' + \frac{g}{\ell}\sin\theta = 0.$$

This is a non-linear, second order, scalar equation. It can*not* be solved by a finite, explicit formula using classical functions.

## Some terminology

**Definition** (imprecise and temporary).

- A differential equation is an equation of the form

$$H(t, X(t), X'(t), \dots, X^{(n)}(t)) = 0$$

  where $H$ is a continuous function.

- It has order $n$ if the highest derivative appearing in the equation is $X^{(n)}$.

- It is explicit if it can be written in the simpler form

$$X^{(n)}(t) = G(t, X(t), X'(t), \dots, X^{(n-1)}(t))$$

- It is linear if $G$ is linear in $X(t), \dots, X^{(n-1)}(t)$; this is bad terminology as one should say affine.

- It is scalar if $X(t)$ is real- or complex-valued, vector if $X(t)$ takes its values in some other normed vector space.

**Definition** (imprecise and temporary). A solution of a differential equation consists of some interval $I \subseteq \mathbb{R}$ and some continuously differentiable function $X$ defined on $I$, subject to the relation.

**Remarks.**

- Most of the time in the non-linear case, $I$ is proper in $\mathbb{R}$; this does not happen in the linear case (where "every solution is maximal").

- Most of the time in the non-linear case, one cannot solve explicitly (i.e. there is no finite-length formula using usual functions; inventing new names, or using a series expansion, can be a possibility); even in the linear case we shall see the phenomenon.

- As a conclusion, *to study is not the same thing as to solve.* To study is to discuss existence, uniqueness, and "qualitative properties" of solutions, even when one cannot write an explicit formula for them.

# Chapter I: Linear Differential Equations

## I.1. Scalar, linear, order 1 equations

**Remark.** Things are too good to last, since here we have:

- an existence and uniqueness result (as a special case of the Cauchy-Lipschitz theorem), which will extend to any *linear* differential equation and even a wide class of non-linear equations, but not all;
- underlying algebraic structures: vector spaces and affine spaces; this will extend to any *linear* differential equation, but will not survive in the non-linear world;
- globality of solutions: a problem posed on $I$ will have a solution defined on $I$; this will extend to any *linear* differential equation, but will not survive in the non-linear world;
- explicit formulas for solutions: this will extend to *linear* equations *with constant coefficients, but not beyond.*

### I.1.1. The "homogeneous" case

**Theorem.** Let $I \subseteq \mathbb{R}$ be an open subinterval and $a : I \to \mathbb{R}$ be a continuous map; also let $t_0 \in I$ and $x_0 \in \mathbb{R}$.

Consider the differential equation $x'(t) = a(t)x(t)$ with initial condition $x(0) = x_0$. Then there is a unique solution; it is defined on $I$, and is given by a useful formula:

$$x(t) = x_0 \cdot e^{\int_{t_0}^t a(s)ds}.$$

**Proof**. A "there exists a unique statement" actually consists of two claims: one on existence, one on uniqueness.

- Existence.

  Let $E(t) = e^{\int_{t_0}^t a(s)ds}$; observe that this map is defined on $I$, and differentiable with derivative:

$$E'(t) = a(t)E(t);$$

moreover, $E(t_0) = e^0 = 1$.

As a consequence, the map $x(t) = x_0 E(t)$ is defined on $I$, differentiable with derivative $x'(t) = x_0 a(t)E(t) = a(t)x(t)$, and satisfies $x(t_0) = x_0$: it is a solution.

- Uniqueness.

  Let $y(t)$ be another solution; for the future's sake we take into account the possibility that $y(t)$ be only defined on some subinterval $J \subseteq I$; this subtlety may certainly be omitted at first reading.

  Consider function:

  $$z(t) = e^{-\int_{t_0}^t a(s)ds} \cdot y(t) = \frac{1}{E(t)} y(t) = (E(t))^{-1} y(t)$$

  (this is however *not* equal to $E(-t)y(t)$, a function of no interest). Clearly $z$ is differentiable on $J$ with derivative:

  $$\begin{aligned} z'(t) &= -a(t)e^{-\int_{t_0}^t a(s)ds} \cdot y(t) + e^{-\int_{t_0}^t a(s)ds} \cdot y'(t) \\ &= -a(t)e^{-\int_{t_0}^t a(s)ds} \cdot y(t) + e^{-\int_{t_0}^t a(s)ds} \cdot a(t)y(t) \\ &= 0, \end{aligned}$$

  so $z$ is a constant function on $J$; in particular, $z(t) = z(t_0) = y(t_0) = x_0$. Then, always on $J$:

  $$y(t) = E(t)z(t) = E(t)x_0 = x(t),$$

  so $y$ coincides with $x$ (at least, where $y$ is defined): proving uniqueness.

$\square$

### I.1.2. The general case

**<u>Theorem</u>.** Let $I \subseteq \mathbb{R}$ be an open subinterval and $a, b : I \to \mathbb{R}$ be continuous maps; also let $t_0 \in I$ and $x_0 \in \mathbb{R}$.

Consider the differential equation $x'(t) = a(t)x(t) + b(t)$ with initial condition $x(0) = x_0$. Then there is a unique solution; it is defined on $I$, and there is a useless formula.

**Proof.**

- Existence.

Use again the integral factor $E(t) = e^{\int_{t_0}^{t} a(s)ds}$; now let:

$$x(t) = E(t) \cdot \left( x_0 + \int_{t_0}^{t} b(s)(E(s))^{-1} ds \right)$$

(The reader had been warned: the formula is completely useless; what matters is the method to get it, a topic to which we shall return.)

Clearly $x$ is defined on $I$, differentiable, with derivative:

$$\begin{aligned}
x'(t) &= E'(t) \cdot \left( x_0 + \int_{t_0}^{t} b(s)(E(s))^{-1} ds \right) + E(t) \cdot b(t)(E(t))^{-1} \\
&= a(t)E(t) \cdot \left( x_0 + \int_{t_0}^{t} b(s)(E(s))^{-1} ds \right) + b(t) \\
&= a(t)x(t) + b(t);
\end{aligned}$$

moreover, $x(t_0) = E(t_0) \cdot (x_0 + 0) = 1 \cdot x_0 = x_0$, so $x$ is a solution.

- Uniqueness.

  Notice that if $y_1, y_2$ are two solutions with the initial condition $y_1(t_0) = y_2(t_0) = x_0$, then $z = y_1 - y_2$ is a solution of the easier problem:

  $z'(t) = a(t)z(t)$ with initial condition $z(t_0) = 0$.

  But the constant map 0 is certainly another solution of the other problem. By the "homogeneous" casee we know that there is uniqueness in the latter. Hence $z = 0$, meaning $y_1 = y_2$. □

**Remark.** To be fully rigorous, what we proved is that $z = 0$ *where it is defined*, i.e. if $y_1$ is defined on $J_1$ and $y_2$ on $J_2$, then we have $y_1 = y_2$ on $J_1 \cap J_2$ (which contains $t_0$). Since the solution $x$ we constructed was defined on all of $I$, this means that both $y_1$ and $y_2$ appear as restrictions of $x$: we have uniqueness in this sense.

This remark can be omitted at first reading, but is typical of our future non-linear arguments.

- **A practical approach**

The formula given in the last theorem is completely useless; do not even bother to try learning it. Here is the method:

- Suppose we wish to solve an equation $(\mathcal{E}) : x'(t) = a(t)x(t) + b(t)$, possibly with initial condition $x(t_0) = x_0$.
- If there was a given initial condition, first forget about it.
- All solutions of the simpler, "homogeneous" equation $(\mathcal{E}_H)$ : $x'(t) = a(t)x(t)$ have the form $\lambda \cdot x_H(t)$, where $\lambda \in \mathbb{R}$ and $x_H$ is any non-zero solution of $(\mathcal{E}_H)$ (which is easily found using integration and exponentiation).
- To find one solution of $(\mathcal{E})$ we look for it in the form $\lambda(t)x_H(t)$.

For $x_S(t) = \lambda(t) \cdot x_H(t)$ to be a solution of $(\mathcal{E})$ one needs:

$$x'_S(t) = \lambda'(t)x_H(t) + \lambda(t)x'_H(t) = \lambda'(t)x_H(t) + \lambda(t)a(t)x_H(t)$$

to be equal to:

$$a(t)x_S(t) + b(t) = a(t)\lambda(t)x_H(t) + b(t).$$

So the condition reduces to $\lambda'(t)x_H(t) = b(t)$, which in practice easily gives $\lambda(t)$.

- Now we have one solution $x_S(t)$ of $(\mathcal{E})$, we see that all solutions of $(\mathcal{E})$ are of the form $x_S(t) + \lambda x_H(t)$, where $\lambda$ is a constant again.
- If an initial condition was given, we finally solve for $\lambda$ in $x_S(t_0)+\lambda x_H(t_0) = x_0$.

The method is called variation of parameters because the parameter $\lambda$ has become a function. Here is a concrete example.

**Examples.**

- On $I = \mathbb{R}_{>0}$, let us solve $x'(t) = \frac{1}{t}x(t) + t^2$ with initial condition $x(2) = 10$.

  We first solve $x'(t) = \frac{1}{t}x(t)$. As we know, solutions are of the form $\lambda \exp\left(\int_1^t \frac{1}{s}ds\right) = \lambda \exp(\log t) = \lambda t$.

  We now look for one special solution of the original equation in the form $x_S(t) = \lambda(t) \cdot t$. For this to be a solution one needs $\lambda'(t) \cdot t + \lambda(t) = \frac{1}{t}\lambda(t) \cdot t + t^2$, i.e. $t\lambda'(t) = t^2$, and finally $\lambda'(t) = t$. So we let $\lambda(t) = \frac{1}{2}t^2$; since the above is revertible, $x_S(t) = \frac{1}{2}t^3$ is a special solution.

  Put together, all solutions of the equation are of the form $\lambda t + \frac{1}{2}t^3$. There is only one satisfying the initial condition $x(t_0) = x_0$; this requires $2\lambda + 4 = 10$, so $\lambda = 3$.

  As a conclusion, the only solution of the equation with initial condition is $x(t) = 3t + \frac{1}{2}t^3$.

- Sometimes finding one solution of $(\mathcal{E})$ is immediate and does not require varying the parameter, as in the previous and also following cases.

  On $I = \mathbb{R}$ consider the equation:

  $$(\mathcal{E}): \quad x'(t) = -tx(t) + (t^2 + 1).$$

  Then equation $(\mathcal{E}_H)$ is $x'(t) = -tx(t)$, which solves into:

  $$x_H(t) = \exp(-tdt) = \exp\left(-\frac{t^2}{2} + c\right) = \lambda e^{-\frac{t^2}{2}}$$

  On the other hand $x_S(t) = t$ is clearly a solution of $(\mathcal{E})$.

### I.1.3. The Algebraist speaks

Forget initial conditions and focus only on the differential equations:

$$\begin{aligned} (\mathcal{E}_H): &\quad x'(t) = a(t)x(t); \\ (\mathcal{E}): &\quad x'(t) = a(t)x(t) + b(t). \end{aligned}$$

Here is the algebraic approach, formulas left aside.

**<u>Theorem</u>.**

- The set $S_H$ of solutions of $(\mathcal{E}_H)$ is a 1-dimensional vector space; for any $t_0 \in I$, the <u>evaluation</u> map:

$$\begin{array}{rccc} ev_{t_0}: & S_H & \to & \mathbb{R} \\ & x_H & \mapsto & x_H(t_0) \end{array}$$

  is a linear isomorphism.
- The set $S$ of solutions of $(\mathcal{E})$ is a 1-dimensional affine space directed by $S$, meaning that for any $x_S \in S$, one has $S = x_S + S_H$.

**Remark.** Informal rephrasing:

- all solutions of $(\mathcal{E}_H)$ are collinear;
- general solution of $(\mathcal{E})$ = one special solution of $(\mathcal{E})$ + general solution of $(\mathcal{E}_H)$.

However proper mathematicians prefer to think in geometric terms.

## I.2. Higher order linear equations

**Remark.** Higher order linear equations still retain:

- an existence and uniqueness result (as a special case of the Cauchy-Lipschitz theorem);
- vector spaces and affine spaces as underlying algebraic structures;
- globality of solutions: a problem posed on $I$ has a solution defined on $I$;
- explicit formulas for solutions *only in the case of constant coefficients.*

### I.2.1. Scalar linear equations with variable coefficients

The following is a special case of the Cauchy-Lipschitz theorem.

**<u>Theorem</u>.** Let $a_1, a_0, b : I \to \mathbb{R}$ be continuous maps; also let $t_0 \in I$ and $x_0, x_1 \in \mathbb{R}$.

Consider the differential equation $x''(t) = a_1(t)x'(t) + a_0(t)x(t) + b(t)$ with initial conditions $x(t_0) = x_0$ and $x'(t_0) = x_1$. Then there is a unique solution; it is defined on $I$... *but there is no formula.*

Return to the two relevant equations:

$$\begin{aligned} (\mathcal{E}): &\quad x''(t) = a_1(t)x'(t) + a_0(t)x(t) + b(t); \\ (\mathcal{E}_H): &\quad x''(t) = a_1(t)x'(t) + a_0(t)x(t) \end{aligned}$$

Here is a nice rephrasement of the above.

**Theorem.**

- The set $S_H$ of solutions of $(\mathcal{E}_H)$ is a 2-dimensional vector space; for any $t_0 \in I$, the evaluation map:

$$\begin{array}{rccc} ev_{t_0}: & S_H & \to & \mathbb{R}^2 \\ & x_H & \mapsto & (x_H(t_0), x'_H(t_0)) \end{array}$$

  is a linear isomorphism.

- The set $S$ of solutions of $(\mathcal{E})$ is a 2-dimensional affine space directed by $S$, meaning that for any $x_S \in S$, one has $S = x_S + S_H$.

## I.2.2. Scalar linear equations with constant coefficients

We briefly recall how to deal with equations of the form:

$$(\mathcal{E}_H): \quad x''(t) = a_1 x'(t) + a_0 x(t)$$

where $a_0, a_1$ are *constant coefficients.*

- Introduce the polynomial equation $\lambda^2 = a_1\lambda + a_0$.
- Solve it (in $\mathbb{C}$), finding complex roots $\lambda_1, \lambda_2$.
- If $\lambda_1 \neq \lambda_2$, then all complex-valued solutions are of the form $c_1e^{\lambda_1 t} + c_2e^{\lambda_2 t}$.
- If $\lambda_1 = \lambda_2$ (so let us simply call it $\lambda$), then all complex-valued solutions are of the form $c_1e^{\lambda t} + c_2te^{\lambda t}$.

This was a clumsy way of stating the following.

**Theorem.** In the assumptions and notation of the section:

- The set $S_H$ of solutions of $(\mathcal{E}_H)$ is a 2-dimensional vector space;
- if $\lambda_1 \neq \lambda_2$, then $(e^{\lambda_1 t}, e^{\lambda_2 t})$ is a basis of $S_H$;
- if $\lambda_1 = \lambda_2$ (so let us simply call it $\lambda$), then $(e^{\lambda t}, te^{\lambda_t})$ is a basis of $S_H$.

More generally, to solve

$$x^{(n)}(t) = a_{n-1}x^{(n-1)}(t) + \cdots + a_1 x'(t) + a_0 x(t),$$

introduce the polynomial $\lambda^n = a_{n-1}\lambda^{n-1} + \cdots + a_1\lambda + a_0$ and solve it in $\mathbb{C}$, finding roots $\lambda_i$ with multiplicity $d_i$.

Then all solutions are linear combinations of the various $t^j e^{\lambda_i t}$ for $j < d_i$.

**Remarks.**

- The good student should be quite dissatisfied with this recipe relying on no proof: this seems to work by miracle, without a clear understanding of what is going on (we will return to this).
- The polynomial involved is called the characteristic equation for a reason which will become obvious once we put the problem into its proper setting: linear algebra.
- In the case of *variable* coefficients, there is a good general theory but no formulas.

### I.2.3. Reduction to equations of order 1

Do you remember the matrix trick from the introduction? We return to it and prove something actually more general. This will be our first incursion into non-linear theory.

**Proposition.** Any differential equation of order $n$ with unknown functions taking values in $\mathbb{R}^d$ can be rewritten as a differential equation *of order* 1 with values in $\mathbb{R}^{nd}$.

In particular, any scalar equation of order $n$ rewrites as a vector equation *of order* 1.

**Proof**. We first do the proof in the scalar case (i.e. when $d = 1$); then we shall inspect our argument and realise it remains valid for any $d$.

The general form of a scalar differential equation of order $n$ (not necessarily linear) is:

$$(\mathcal{E}): \qquad x^{(n)}(t) = F(t, x(t), x'(t), \dots, x^{(n-1)}(t)),$$

where $F(t, a_0, \dots, a_{n-1})$ is a continuous map defined on some subset of $I \times \mathbb{R} \times \cdots \times \mathbb{R}$ and taking values in $\mathbb{R}$. All this may seem a bit imprecise for the moment but the chapter on non-linear equations will be extremely rigorous on the topic.

Introduce the function:

$$\vec{\Phi}(t, v_0, \dots, v_{n-1}) = (v_1, \dots, v_{n-1}, F(t, v_0, \dots, v_{n-1})),$$

a continuous map defined on the same subset of $I \times \mathbb{R} \times \cdots \times \mathbb{R}$ as $F$, but taking values in $\mathbb{R}^n$.

Now consider the equation:

$$(\vec{\mathcal{E}}) : \qquad \vec{X}'(t) = \vec{\Phi}(t, \vec{X}(t)) \qquad \text{for } t \in I$$

We claim that $y : I \to \mathbb{R}$ is a solution of $(\mathcal{E})$ iff the function:

$$\begin{array}{rrcc} \vec{Y} : & I & \to & \mathbb{R}^n \\ & t & \mapsto & \begin{pmatrix} y(t) \\ y'(t) \\ \vdots \\ y^{(n-1)}(t) \end{pmatrix} \end{array}$$

is a solution of $(\vec{\mathcal{E}})$.

- Suppose that $y$ is a solution of $(\mathcal{E})$ and let $\vec{Y}$ as above. Then:

$$\vec{Y}'(t) = \begin{pmatrix} y'(t) \\ \vdots \\ y^{(n-1)}(t) \\ y^{(n)}(t) \end{pmatrix} = \begin{pmatrix} y'(t) \\ \vdots \\ y^{(n-1)}(t) \\ F(t, y(t), y'(t), \ldots, y^{(n-1)}(t)) \end{pmatrix} = \vec{\Phi}\left(t, \vec{Y}(t)\right)$$

- The converse is similar.

For the general case ($d > 1$), read the proof again, replacing every occurence of $\mathbb{R}$ by $\mathbb{R}^d$: the argument is still valid. □

**Corollary.** Any scalar linear equation of order $n$ rewrites as a vector linear equation of order 1.

**Proof**. A scalar linear equation has the special form:

$$(\mathcal{E}) : \qquad x^{(n)}(t) = a_0(t)x(t) + a_1(t)x'(t) + \cdots + a_{n-1}(t)x^{(n-1)}(t) + b(t),$$

meaning that here:

$$F(t, v_0, \ldots, v_{n-1}) = a_0(t)v_0 + a_1(t)v_1 + \cdots + a_{n-1}(t)v_{n-1} + b(t).$$

Consequently,

$$\begin{aligned}\vec{\Phi}(t, v_0, \dots, v_{n-1}) &= (v_1, \dots, v_{n-1}, a_0(t)v_0 + \cdots + a_{n-1}(t)v_{n-1} + b(t)) \\ &= \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 1 \\ a_0(t) & a_1(t) & \dots & \dots & a_{n-1}(t) \end{pmatrix} \cdot \begin{pmatrix} v_0 \\ \vdots \\ \vdots \\ v_{n-1} \end{pmatrix} + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ b(t) \end{pmatrix} \\ &= A(t) \cdot \vec{V} + B(t)\end{aligned}$$

in obvious notation. □

**Remark.** A similar argument shows that any system of coupled equations rewrites as one (big) first-order equation; if the system is linear, so is the resulting equation.

**Example.** We give a non-linear example. Return to the Lotka-Volterra system:

$$\begin{cases} p_1'(t) &= p_1(t) \cdot (a - bp_2(t)) \\ p_2'(t) &= p_2(t) \cdot (-c + dp_1(t)) \end{cases}.$$

Introduce the function:

$$G(t, v_1, v_2) = (v_1 \cdot (a - bv_2), v_2 \cdot (-c + dv_1)).$$

Then $(p_1, p_2)$ is a solution of the coupled system iff $P(t) = \begin{pmatrix} p_1(t) \\ p_2(t) \end{pmatrix}$ is a solution of $X'(t) = G(t, X(t))$.

# I.3. Matrix exponential

Now comes something different.

We will try to generalise the scalar, order 1 *linear* formulas to vector, order $n$, *linear* differential equations. The formula giving the solution to a scalar, order 1, linear Cauchy problem was, in our usual notation:

$$x(t) = e^{\int_{t_0}^t a(s)ds} \cdot x_0$$

Be careful that the generalisation of this formula to the vector case (where $A(s)$ is a matrix) will *not* be correct. It will be only in the case of a *constant* matrix $A$. But anyway, we must first understand what it means to take the exponential of a matrix.

## I.3.1. The definition

**Theorem** (and **Definition**). Let $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$ and $A \in M_d(\mathbb{K})$. Then the series:

$$\sum_{k \geqslant 0} \frac{1}{k!} A^k$$

converges to a limit, called the exponential of $A$ and denoted $\exp(A)$.

**Remark.** From now on we denote the matrix exponential by exp, and reserve notation $e^z$ for the *real/complex* exponential.

**Proof.**

**Step 1.** A fact from matrix algebra: there is a norm $|||\cdot|||$ on $M_d(\mathbb{K})$ such that for any $M, N \in M_d(\mathbb{K})$ one has:

$$|||M \cdot N||| \leqslant |||M||| \cdot |||N|||.$$

We say that $|||\cdot|||$ is sub-multiplicative.

**Proof.** Fix any norm $\|\cdot\|$ on $\mathbb{K}^d$ (for instance, the usual norm). We let $S = \{v \in \mathbb{K}^d : \|v\| = 1\}$ be the associated sphere. For $M \in M_d(\mathbb{K})$ we let:

$$|||M||| = \sup_{v \in S} \|Mv\|,$$

and we claim that $|||\cdot|||$ meets the requirements. We therefore have several properties to check.

- $|||M|||$ is well-defined, i.e. the sup takes finite values. Indeed, the map:

$$\begin{array}{rccc} f: & S & \to & \mathbb{R}_{\geqslant 0} \\ & v & \mapsto & \|Mv\| \end{array}$$

  is the composition of $\|\cdot\|$ and with multiplication by $M$. The first is continuous as it is 1-Lipschitz; the second is continuous since it is linear. As a conclusion, $f$ is continuous. Since $S$ is closed and bounded and $\mathbb{K}^d$ it is compact; therefore $f$ is bounded on $S$ and its supremum is finite (it even is a maximum).

- $|||\cdot|||$ defines a norm. Three properties to check briefly.
  - $|||\cdot|||$ takes non-negative values and is zero only on the null matrix. Indeed, if $|||M||| = 0$ then $S \subseteq \ker M$. Now $S$ spans the whole space $\mathbb{K}^d$ and $\ker M$ is a vector subspace; this shows $\ker M = \mathbb{K}^d$, and $M = 0$.
  - $|||\cdot|||$ is positively homogeneous.
    Indeed, let $M \in M_d(\mathbb{K})$ and $\lambda \in \mathbb{K}$; then for any $v \in S$ one has:

$$\|(\lambda M) \cdot v\| = |\lambda| \cdot \|\lambda M \cdot v\|$$

    so taking the supremum on both sides, $|||\lambda M||| = |\lambda| |||M|||$.
  - $|||\cdot|||$ satisfies the triangle inequality.

Indeed, let $M, N \in M_d(\mathbb{K})$. For $v \in S$, one has:

$$\|(M+N)v\| = \|Mv + Nv\| \leqslant \|Mv\| + \|Nv\| \leqslant |||M||| + |||N|||,$$

so taking the supremum, $|||M+N||| \leqslant |||M||| + |||N|||$.

This proves that $|||\cdot|||$ is a norm on $M_d(\mathbb{K})$.

- $|||\cdot|||$ is sub-multiplicative.

  Indeed, let $M, N \in M_d(\mathbb{K})$. Let $v \in \mathbb{K}^d$. If $Nv = 0$ then $\|MNv\| = \|0\| \leqslant |||M||| \cdot |||N|||$, this case is not very interesting. Otherwise let $w = \frac{1}{\|Nv\|} Nv \in S$, so:

  $$\|MNv\| = \|M \cdot \|Nv\| w\| = \|Nv\| \cdot \|Mw\| \leqslant |||N||| \cdot |||M|||$$

  Taking the supremum on $v \in S$, we find $|||MN||| \leqslant |||M||| \cdot |||N|||$, as desired.

$\Diamond$

A norm of the norm $|||A||| = \sup_{\|v\|=1} \|Av\|$, for $\|\cdot\|$ a vector norm, is often called an operator norm.

**Step 2.** The series $\sum_{k=0}^{n} \frac{1}{k!} A^k$ converges.

**Proof**. As a consequence of sub-multiplicativity, one has by induction:

$$|||A^n||| \leqslant |||A|||^n$$

Hence:

$$\left|\left|\left| \sum_{k=0}^{n} \frac{1}{k!} A^k \right|\right|\right| \leqslant \sum_{k=0}^{n} \frac{1}{k!} |||A^k||| \leqslant \sum_{k=0}^{n} \frac{1}{k!} |||A|||^k \leqslant e^{|||A|||},$$

which suggests that the series $\sum_{k=0}^{n} \frac{1}{k!} A^k$ should converge. This follows from some topological facts.

**Definition.** Let $(X, d)$ be a metric space.

- A sequence $(u_n) \in X^{\mathbb{N}}$ is a Cauchy sequence if:

  $$\forall \varepsilon > 0 \exists N \in \mathbb{N} \forall q \geqslant p \geqslant N, \quad d(u_p, u_q) < \varepsilon$$

- $(X, d)$ is a complete metric space if all Cauchy sequences are convergent.

**Fact.** For any norm $\nu$, the metric space $(M_d(\mathbb{K}), \nu)$ is complete.

This is easy to believe since a sequence $(M_n)$ converges iff each sequence

$(m_n^{i,j})$ converges, where $(m_n^{i,j})$ are the coefficients of the matrix $M_n$.

**Definition.** Let $(E, \nu)$ be a normed vector space. A series $S_n = \sum_{k=0}^{n} u_k$ is normally convergent if the *real* series $\sum_{k=0}^{n} u_k \nu(u_k)$ is convergent.

**Fact.** If $(E, \nu)$ is a normed vector space which is complete, then every normally convergent series is convergent.

Let us put the pieces together. We have found one norm $|||\cdot|||$ on $M_d(\mathbb{K})$ such that the series $\sum \frac{1}{k!} A^k$ is normally convergent. By completeness of the space, it is therefore convergent. $\Diamond$

So $\exp(A)$ is well-defined. $\square$

### I.3.2. Exponential of a diagonalisable matrix

The involved proof of existence is absolutely not efficient in practice: computing a given matrix' exponential remains a mystery. *It is certainly not componentwise exponentiation.* Let us start with basic examples.

**Examples.**

- Start with the null matrix. Then clearly:

$$\exp(0) = \frac{1}{0!}0^0 + \frac{1}{1!}0^1 + \frac{1}{2!}0^2 + \dots$$

  The first term is the identity matrix $I$ (as will be for any matrix); the second is 0, and actually any next term is a positive power of 0, so 0. Consequently $\exp(0) = I$.

- Let us compute $\exp(I)$. Now:

$$\exp(I) = \frac{1}{0!}I + \frac{1}{1!}I + \frac{1}{2!}I^2 + \dots,$$

  so clearly $\exp(I) = e \cdot I$. A more interesting way to view this is by writing the diagonal explicitly:

$$\begin{aligned}\exp\begin{pmatrix}1 & & \\ & \ddots & \\ & & 1\end{pmatrix} &= \begin{pmatrix}\frac{1}{0!}+\frac{1}{1!}+\frac{1}{2!}+\dots & & \\ & \ddots & \\ & & \frac{1}{0!}+\frac{1}{1!}+\frac{1}{2!}+\dots\end{pmatrix} \\ &= \begin{pmatrix}e & & \\ & \ddots & \\ & & e\end{pmatrix}\end{aligned}$$

- This generalises neatly: suppose we have a diagonal matrix. Then:

$$\exp\begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_n\end{pmatrix} = \begin{pmatrix}e^{\lambda_1} & & \\ & \ddots & \\ & & e^{\lambda_n}\end{pmatrix}$$

In the case of a diagonal matrix we can compute the exponential quickly: it then is componentwise exponentiation on the diagonal (this will not be true if there are non-diagonal terms). We can extend to diagonalisable matrices thanks to the following.

**Proposition** (conjugacy property)**.** Suppose $A, P \in M_d(\mathbb{K})$ and $P$ is invertible. Then $\exp(PAP^{-1}) = P\exp(A)P^{-1}$.

**Proof**. For any matrix $M$ and integer $n$, let:

$$S_n(M) = \sum_{k=0}^{n} \frac{1}{k!} M^k.$$

As we know, for any $M \in M_d(\mathbb{K})$, one has $S_n(M) \to \exp(M)$. Also introduce the map:

$$\begin{array}{rccc} f: & M_d(\mathbb{K}) & \to & M_d(\mathbb{K}) \\ & M & \mapsto & PMP^{-1} \end{array}$$

which is called conjugation by $P$. Since $f$ is linear, it is continuous. Now observe how for any matrix $M$, one has $f(S_n(M)) = S_n(f(M))$. Hence on the one hand:

$$f(S_n(A)) = S_n(f(A)) \to \exp(f(A)) = \exp(PAP^{-1});$$

on the other hand, since $f$ is continuous at $\exp(A)$ and $S_n(A) \to \exp(A)$:

$$f(S_n(A)) \to f(\exp A) = P\exp(A)P^{-1}.$$

The limit being unique, we are done. □

### I.3.3. Matrix exponential: the general case

**Proposition** (commutation property)**.** Let $A, B \in M_d(\mathbb{K})$ be such that $AB = BA$ (one says that $A$ and $B$ commute). Then $\exp(A+B) = \exp(A)\cdot\exp(B) = \exp(B)\cdot\exp(A)$.

**Proof**. This is a standard manipulation of absolutely (here, normally) converging series and does not say much about matrix exponential nor differential equations. □

**Example.** The assumption is required; although we still cannot compute matrix exponentials in general we can already give a counterexample.

Let $A = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ and $B = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$.

First of all:

$$AB = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \cdot \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}; BA = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix},$$

so the proposition does not apply.

One sees that $A^2 = I$, so:

$$\exp(A) = I + A + \frac{1}{2}I + \frac{1}{6}A + \cdots = \begin{pmatrix} e & \\ & e^{-1} \end{pmatrix},$$

while $B^2 = 0$, so:

$$\exp B = I + B + 0 + \cdots = I + B = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$

and therefore:

$$\exp(A) \cdot \exp(B) = \begin{pmatrix} e & \\ & e^{-1} \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} e & e \\ & e^{-1} \end{pmatrix}.$$

Now in order to compute $\exp(A + B)$, we see that:

$$A + B = \begin{pmatrix} 1 & 1 \\ & -1 \end{pmatrix},$$

and:

$$(A + B)^2 = \begin{pmatrix} 1 & 1 \\ & -1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 \\ & -1 \end{pmatrix} = \begin{pmatrix} 1 & \\ & 1 \end{pmatrix},$$

so that:

$$\exp(A + B) = I + (A + B) + \frac{1}{2!}I + \frac{1}{6!}A + B + \cdots = \begin{pmatrix} e & \sum_{k \text{ odd}} \frac{1}{k!} \\ & e^{-1} \end{pmatrix},$$

which differs from $\exp(A) \cdot \exp(B)$.

**Definition.** A matrix $N \in M_d(\mathbb{C})$ is nilpotent if $N^d = 0$.

**Remark.** It is an excellent exercise in linear algebra to prove the equivalences between:

- $N$ is nilpotent;

- there exists $k \in \mathbb{N}$ such that $N^k = 0$;
- 0 is the only eigenvalue of $N$.

**Remark.** Let $N \in M_d(\mathbb{C})$ be a nilpotent matrix. Then:

$$\exp(N) = I + N + \cdots + \frac{1}{(d-1)!} N^{d-1}$$

since the following terms of the series all vanish.

**Proposition** (Dunford decomposition)**.** Let $M \in M_d(\mathbb{C})$. Then there is a unique pair $(D, N)$ with:

- $D$ a diagonalisable matrix,
- $N$ a nilpotent matrix,
- $M = D + N$,
- $DN = ND$ ($N$ and $D$ commute).

We finally get a practical method: to compute $\exp A$, we try to write it in the above form; then $\exp A = \exp D \cdot \exp N$.

**Remarks.**

- Always treat your matrix as a complex matrix.
- Be extremely careful: writing $A = D + N$ is not enough, *one must check that $D$ and $N$ commute.*

**Examples.**

- Let $A = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$; we compute $\exp(A)$.

  Be careful that here the naive decomposition:

  $$A = \begin{pmatrix} 1 & \\ & 2 \end{pmatrix} + \begin{pmatrix} 0 & 1 \\ & 0 \end{pmatrix}$$

  is of no use, since the terms do not commute.

  On the other hand the matrix is diagonalisable at first sight:

  $$A \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad \text{and} \qquad A \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 2 \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

  so

  $$\left( \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right)$$

is an eigenbasis. So we introduce the coordinate change matrix

$$P = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$

for which:

$$A = P \begin{pmatrix} 1 & \\ & 2 \end{pmatrix} P^{-1}.$$

Now clearly:

$$\begin{aligned}
\exp(A) &= \exp\left(P \begin{pmatrix} 1 & \\ & 2 \end{pmatrix} P^{-1}\right) \\
&= P \exp \begin{pmatrix} 1 & \\ & 2 \end{pmatrix} P^{-1} \\
&= \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} e & \\ & e^2 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \\
&= \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} e & -e \\ 0 & e^2 \end{pmatrix} \\
&= \begin{pmatrix} e & e - e^2 \\ 0 & e^2 \end{pmatrix}
\end{aligned}$$

- Now let $B = \begin{pmatrix} 1/2 & 1/2 \\ -1/2 & 3/2 \end{pmatrix}$; let us find $\exp(B)$.

  The determinant is 1 and trace is 2; clearly $B$ has the sole eigenvalue 1 with algebraic multiplicity 2. Let us determine the associated eigenspace:

  $$E_1(B) = \ker(B - I) = \ker \begin{pmatrix} -1/2 & 1/2 \\ -1/2 & 1/2 \end{pmatrix} = \ker \begin{pmatrix} 1 & -1 \end{pmatrix} = \operatorname{Span} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

  so $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ is an eigenvector. In dimension 2 (for reasons belonging to your linear algebra class and which I had not time to explain) it is safe to extend to any basis; the resulting coordinate change matrix will trigonalise $B$.

  So we let:

  $$Q = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix};$$

  it can be checked that

  $$Q^{-1} B Q = \begin{pmatrix} 1 & -1/2 \\ & 1 \end{pmatrix}.$$

Now observe how:

$$\begin{aligned}
\exp\begin{pmatrix} 1 & -1/2 \\ & 1 \end{pmatrix} &= \exp\left(\begin{pmatrix} 1 & \\ & 1 \end{pmatrix} + \begin{pmatrix} 0 & -1/2 \\ & 0 \end{pmatrix}\right) \\
&= \exp\begin{pmatrix} 1 & \\ & 1 \end{pmatrix} \cdot \exp\begin{pmatrix} 0 & -1/2 \\ & 0 \end{pmatrix} \\
&= \begin{pmatrix} e & \\ & e \end{pmatrix} \cdot \begin{pmatrix} 1 & -1/2 \\ & 1 \end{pmatrix} \\
&= \begin{pmatrix} e & -1/2e \\ & e \end{pmatrix},
\end{aligned}$$

so we finally find:

$$\begin{aligned}
\exp(B) &= \exp\left(Q \begin{pmatrix} 1 & -1/2 \\ & 1 \end{pmatrix} Q^{-1}\right) \\
&= Q \cdot \exp\begin{pmatrix} 1 & -1/2 \\ & 1 \end{pmatrix} \cdot Q^{-1} \\
&= \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} e & -1/2e \\ & e \end{pmatrix} \cdot \begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix} \\
&= \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} -1/2e & 3/2e \\ e & -e \end{pmatrix} \\
&= \begin{pmatrix} e/2 & e/2 \\ -e/2 & 3/2e \end{pmatrix}
\end{aligned}$$

### I.3.4. Similarity classes for 2-dimensional complex matrices

Recall that two matrices $A, B$ are similar if there is an invertible matrix $P$ with $A = PBP^{-1}$ (or equivalently, there is invertible $Q$ with $A = Q^{-1}BQ$). The following proposition classifies complex $2 \times 2$ matrices up to similarity.

**Proposition.** Let $A \in M_2(\mathbb{C})$. Then there are two cases:

- either $A$ is similar to $\begin{pmatrix} \lambda & \\ & \mu \end{pmatrix}$, possibly with $\lambda = \mu$,
- or $A$ is similar to $\begin{pmatrix} \lambda & 1 \\ & \lambda \end{pmatrix}$, possibly with $\lambda = 0$.

**Proof**. The first case exactly means that $A$ is diagonalisable. So suppose that $A$ is *not* diagonalisable. Then it must have only one eigenvalue, say $\lambda$ (possibly 0), and $\dim E_\lambda(A) = 1$.

Let $v_1 \in E_\lambda(A) \setminus \{0\}$ (i.e. an eigenvector) and $v_2 \notin \mathbb{C}v_1$. Then $(v_1, v_2)$ forms a basis; moreover $(A - \lambda I)^2 = 0$, so:

$$(A - \lambda I)v_2 \in \ker(A - \lambda I) = E_\lambda(A) = \mathbb{C}v_1,$$

meaning that there is $z \in \mathbb{C}$ with $Av_2 = \lambda v_2 + zv_1$. However $v_2$ is *not* an eigenvector, so $z \neq 0$.

Let $v_1' = zv_1$; since $z \neq 0$, the pair $(v_1', v_2)$ still forms a basis. But one has $Av_1' = \lambda v_1'$ and:

$$Av_2 = \lambda v_2 + zv_1 = \lambda v_2 + v_1',$$

so the coordinate change matrix from the standard basis to $(v_1', v_2)$ brings $A$ to $\begin{pmatrix} \lambda & 1 \\ & \lambda \end{pmatrix}$. □

# I.4. Vector linear differential equations with constant coefficients

**Remark.** What follows applies only to linear equations *with constant coefficients.*

Let us finally apply our new tool (the matrix exponential) to the resolution of linear differential equations with constant coefficients.

## I.4.1. Differentiating the exponential

**Lemma.** For $A \in M_d(\mathbb{K})$, the map $t \mapsto \exp(tA)$ is differentiable on $\mathbb{R}$ and:

$$(\exp(tA))' = A \cdot \exp(tA)$$

**Proof**. Let $t_0, h \in \mathbb{R}$. We shall let $h$ go to 0, so we may suppose $|h| < 1$. Observe that $t_0A$ and $hA$ commute, so by the commutation property:

$$\begin{aligned}\frac{1}{h}\left(\exp\left((t_0+h)A\right) - \exp\left(t_0A\right)\right) &= \frac{1}{h}\left(\exp\left(hA\right)\cdot\exp\left(t_0A\right) - \exp\left(t_0A\right)\right) \\ &= \frac{1}{h}(\exp(hA) - I)\exp(t_0A)\end{aligned}$$

Now observe that by definition:

$$\exp(hA) - I \quad = \quad \sum_{k=1}^{\infty}\frac{1}{k!}h^kA^k \quad = \quad hA + \underbrace{\sum_{k=2}^{\infty}\frac{1}{k!}h^kA^k}_{\varepsilon(h)}.$$

The term $\varepsilon(h)$ is however small, as follows:

$$|||\varepsilon(h)||| = \left|\left|\left|\sum_{k=2}^{\infty}\frac{1}{k!}h^kA^k\right|\right|\right| \leqslant h^2\sum_{k=2}^{\infty}\frac{1}{k!}\,|||A|||^k \leqslant h^2e^{|||A|||}$$

In particular:

$$\frac{1}{h}\varepsilon(h) \leqslant h e^{|||A|||} \xrightarrow[h\to 0]{} 0$$

As a conclusion:

$$\begin{aligned}
\frac{1}{h}\left(\exp\left((t_0+h)A\right) - \exp\left(t_0 A\right)\right) &= \frac{1}{h}(\exp(hA) - I)\exp(t_0 A) \\
&= \frac{1}{h}(hA + \varepsilon(h))\exp(t_0 A) \\
&= A\exp(t_0 A) + \frac{1}{h}\varepsilon(h)\exp(t_0 A) \\
&\xrightarrow[h\to 0]{} A\exp(t_0 A),
\end{aligned}$$

which exactly means that $\exp(tA)$ is differentiable at $t_0$, with derivative $A\exp(t_0 A)$. □

**Remarks.**

- A more analytic proof is possible, using a differentiation theorem for uniformly convergent function series.
- It can actually be proved in calculus that the map $\exp : M_d(\mathbb{K}) \to M_d(\mathbb{K})$ is differentiable (as a matter of fact, $C^\infty$). We can even compute the differential at the null matrix:

  $$D_0 \exp = \mathrm{Id},$$

  meaning that for any small matrix $M$, one has $\exp(M) = I + M + o(M)$; this is a slightly stronger statement than the lemma.

  With some pain the differential can be computed at any point, but gives rise to an unpleasant infinite series — something impossible to use in practice.

## I.4.2. Constant coefficients

**Proposition.** Let $I \subseteq \mathbb{R}$ be an open interval, $A \in M_d(\mathbb{R})$ be a matrix, $t_0 \in I$ and $X_0 \in \mathbb{R}^d$. Consider equation:

$$(\vec{\mathcal{E}}) : \qquad X'(t) = A \cdot X(t)$$

with initial condition $X(t_0) = X_0$.

Then there is a unique solution, which is global (i.e. defined over $I$), and we have a useful formula:

$$X(t) = \exp\left(t - t_0)A\right) \cdot X_0$$

**Proof.**

- Existence.

  By the properties of the exponential, the suggested function is differentiable on $I$ with derivative:

  $$X'(t) = A \exp{(t - t_0)A)} \cdot X_0 = AX(t);$$

  moreover $X(t_0) = \exp(0) \cdot X_0 = I \cdot X_0 = X_0$. So the suggested function is a solution.

- Uniqueness.

  Let $Y(t)$ be another solution (defined on some subinterval $J \subseteq I$). Let $Z(t) = \exp(-(t - t_0)A) \cdot Y(t)$, which is again differentiable on $J$ with derivative:

  $$\begin{aligned} Z'(t) &= -A \exp(-(t - t_0)A)Y(t) + \exp(-(t - t_0)A)Y'(t) \\ &= -A \exp(-(t - t_0)A)Y(t) + \exp(-(t - t_0)A)AY(t) \\ &= 0 \end{aligned}$$

  since $Y(t)$ is a solution and $A$ and $\exp(tA)$ commute. So $Z$ is a constant on $J$; hence for $t \in J$ one has $Z(t) = Z(t_0) = \exp(0) \cdot Y(t_0) = X_0$, and:

  $$Y(t) = \exp((t - t_0)A) \cdot Z(t) = \exp((t - t_0)A) \cdot X_0 = X(t).$$

  Therefore function $Y$ appears as the restriction to $J$ of function $X$. □

- **Relationship with the "characteristic method"**

In the case of say $x'' = bx' + cx$, we now have *two* methods.

1. The characteristic equation method: introduce the polynomial $\lambda^2 - b\lambda - c$, with (possibly complex) solutions $\lambda_1, \lambda_2$.

   - If $\lambda_1 \neq \lambda_2$ then all solutions (with no specified initial condition) are of the form:
     $$c_1 e^{\lambda_1 t} + c_2 e^{\lambda_2 t}$$
   - If on the other hand $\lambda_1 = \lambda_2$ (simply denoted $\lambda$) then solutions assume the form:
     $$c_1 e^{\lambda t} + c_2 t e^{\lambda t}$$

   This method avoid linear algebra — and as a result hides the fundamental phenomena.

2. The matrix exponential method: change the problem into solving vector equation
   $$X'(t) = A \cdot X(t)$$
   where $A = \begin{pmatrix} 0 & 1 \\ c & b \end{pmatrix}$ and $X(t) = \begin{pmatrix} x(t) \\ x'(t) \end{pmatrix}$. All solutions have the form $\exp(tA) \cdot C$, where $C \in \mathbb{R}^2$.

We must still explain why both methods will agree.

**Proof**. First notice that the characteristic polynomial of matrix $A$ is:

$$\chi_A(\lambda) = -\lambda(b-\lambda) - c = \lambda^2 - b\lambda - c,$$

which is why we also called it the characteristic polynomial of the equation. Let us keep calling $\lambda_1, \lambda_2$ its complex roots. Now there are two cases.

- Suppose $\lambda_1 \neq \lambda_2$. Then $A$ is diagonalisable: there is invertible $P$ such that:

  $$P^{-1}AP = \begin{pmatrix} \lambda_1 & \\ & \lambda_2 \end{pmatrix}$$

  Then $P(tA)P^{-1}$ is easily computed. Therefore, thanks to the conjugacy property:

  $$\exp(tA) = P \begin{pmatrix} e^{\lambda_1 t} & \\ & e^{\lambda_2 t} \end{pmatrix} P^{-1}$$

  so an arbitrary solution has the form:

  $$X(t) = P \begin{pmatrix} e^{\lambda_1 t} & \\ & e^{\lambda_2 t} \end{pmatrix} P^{-1} \cdot X_1 = P \begin{pmatrix} e^{\lambda_1 t} & \\ & e^{\lambda_2 t} \end{pmatrix} \cdot X_2$$

  Without actually computing it, we see that this yields a vector of the form:

  $$X(t) = \begin{pmatrix} c_1 e^{\lambda_1 t} + c_2 e^{\lambda_2 t} \\ d_1 e^{\lambda_1 t} + d_2 e^{\lambda_2 t} \end{pmatrix},$$

  confirming the old-fashioned method.

- Now suppose $\lambda_1 = \lambda_2$ (denoted $\lambda$). For a general matrix this is no full obstruction to being diagonalisable; but here, it is. For if $A$ were diagonalisable we would have $A = P^{-1}(\lambda I)P$ for some invertible $P$. Now the matrix in the middle commutes to *every* matrix, so we find $A = \lambda I$, which is clearly not the case.

  So $A$ is *not* diagonalisable, and we know that there is invertible $P$ with:

  $$A = P^{-1} \begin{pmatrix} \lambda & 1 \\ & \lambda \end{pmatrix} P$$

  In this case we find:

  $$\exp(tA) = P \begin{pmatrix} e^{\lambda t} & t \\ & e^{\lambda t} \end{pmatrix} P^{-1}$$

  so an arbitrary solution will have the form:

  $$X(t) = P \begin{pmatrix} e^{\lambda t} & t \\ & e^{\lambda t} \end{pmatrix} P^{-1} \cdot X_1 = P \begin{pmatrix} e^{\lambda t} & t \\ & e^{\lambda t} \end{pmatrix} \cdot X_2 = \begin{pmatrix} c_1 e^{\lambda t} + c_2 t e^{\lambda t} \\ d_1 e^{\lambda t} + d_2 t e^{\lambda t} \end{pmatrix}.$$

In either case, both methods yield the same solution space. $\Diamond$

**Remark.** More in general, the linear equation with constant coefficients

$$x^{(n)} = a_{n-1}x^{(n-1)} + \cdots + a_0 x$$

relies on the characteristic polynomial

$$\lambda^n - a_{n-1}\lambda^{n-1} - \cdots - a_0.$$

This is because in matrix form, one has:

$$X'(t) = A \cdot X(t)$$

where:

$$A = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 1 \\ a_0 & a_1 & \dots & \dots & a_{n-1} \end{pmatrix}$$

But the latter has characteristic polynomial $(-1)^n\lambda^n - a_{n-1}\lambda^{n-1} - \cdots - a_0$, as a quick induction and expansion along the first column reveals.

This finally explains the so-called "characteristic equation".

### I.4.3. Non-homogeneous case

**Proposition.** Let $I \subseteq \mathbb{R}$ be an open interval, $A \in M_d(\mathbb{R})$ be a matrix, $B : I \to \mathbb{R}^d$ be a continuous map; also let $t_0 \in I$ and $X_0 \in \mathbb{R}^d$. Consider equation:

$$(\vec{\mathcal{E}}): \qquad X'(t) = A \cdot X(t) + B(t)$$

with initial condition $X(t_0) = X_0$.

Then there is a unique solution, which is global (i.e. defined over $I$), and we have a useless formula.

**Proof**.

- Existence.

  It is easy to check that letting:

  $$X(t) = \exp((t-t_0)A) \cdot \left( X_0 + \int_{t_0}^{t} \exp(-sA)B(s)ds \right)$$

  gives a solution.

- Uniqueness is handled as in the scalar case: if $Y_1, Y_2$ are two solutions (with the initial condition), then $Y_1 - Y_2$ is a solution of $X'(t) = AX(t)$ vanishing at $t_0$, so whereever defined. $\square$

What matters algebraically is the following.

**Theorem.** Let $A \in M_d(\mathbb{R})$ be fixed and $B : I \to \mathbb{R}^d$ be a continuous map. Consider as always equations $(\mathcal{E}) : X'(t) = A \cdot X(t) + B(t)$ and $(\mathcal{E}_H) : X'(t) = A \cdot X(t)$.

- The set $S_H$ of solutions of $(\mathcal{E}_H)$ is a $d$-dimensional vector space; for any $t_0 \in I$, the evaluation map:

$$ev_{t_0} : \begin{array}{ccc} S_H & \to & \mathbb{R}^d \\ x_H & \mapsto & x_H(t_0) \end{array}$$

  is a linear isomorphism.

- The set $S$ of solutions of $(\mathcal{E})$ is a $d$-dimensional affine space directed by $S$, meaning that for any $x_S \in S$, one has $S = x_S + S_H$.

**Remark.** Of course the formula is an instance of the variation of parameters: a special solution must be looked for in the form:

$$x_S(t) = \exp((t - t_0)A) \cdot \Lambda(t),$$

where $\Lambda : I \to \mathbb{R}^d$ is differentiable. This always succeeds (possibly through painful computations if $d$ is large).

Be extremely careful however that in general, this *does not work* for a variable matrix $A(t)$. The naive attempt is simply incorrect; there are no more general formulas here.

# Chapter II: Ordinary Differential Equations

Nature is not linear. Return to our early examples; as simple a model as the ideal pendulum gave rise to a non-linear equation.

The study of non-linear equations is therefore unavoidable. Several warnings must be issued immediately, in contrast to the linear case:

1. there are no formulas; as a matter of fact there is no general method of explicit resolution, and non-linear equations which can be solved by a closed formula are the exception, not the rule;
2. the set of solutions does not bear any algebraic structure (as opposed to the linear case, where it was an affine space);
3. solutions tend not to be global, i.e. tend to be defined on proper subintervals.

## II.1. Terminology and Phenomena

### II.1.1. Terminology

Let us return to the fundamentals.

**Definition.** A differential equation is an equation of the form:

$$(\mathcal{E}): \qquad X'(t) = G(t, X(t)) \quad \text{for } t \in I$$

where:

- $I \subseteq \mathbb{R}$ is a (non-empty) open subinterval;
- $U \subseteq \mathbb{R}^d$ is a (non-empty) open subset;
- $G : I \times U \to \mathbb{R}^d$ is a map *which we will always take to be continuous.*

**Examples.**

- all linear differential equations of course;
- $x'(t) = \sqrt{x(t)}$ is a non-linear equation;

- $\theta''(t) = \sin(\theta(t))$, non-linear again.

It is a good exercise to explicit $I$, $U$, and $G$ in the above to make sure you understand the definition; in particular any differential equation can be converted into one with order 1.

We can also add initial conditions.

**Definition.** An initial condition problem, also known as a Cauchy problem, is given by a differential equation $(\mathcal{E})$ together with: $t_0 \in I$, $X_0 \in U$, and the requirement:
$$X(t_0) = X_0$$

There is a nice theory of Cauchy problems as we shall see, but:

- there are no formulas;
- there are no nice "algebraic" correspondences (eg. linear maps) between solutions and initial conditions;
- solutions tend not to be global.

The latter requires us to review the terminology.

**Definition.** A solution to $(\mathcal{E})$ is a *pair* $(J, X)$, with:

- $J \subseteq I$ a non-empty open subinterval;
- $X : J \to \mathbb{R}^d$ is a differentiable map taking values in $U$;
- $\forall t \in J, \quad X'(t) = G(t, X(t))$.

For a solution to a Cauchy problem we also require $t_0 \in J$ and $X(t_0) = X_0$.

**Definition.** A solution is called global if $J = I$, local otherwise.

This terminology is however far from satisfactory: for instance, if $(J, X)$ is a solution, $\check{J} \subseteq J$ is an open subinterval, and $\check{X} = X_{|\check{J}}$ (the restriction of $X$ to $\check{J}$, then:

- on the one hand, $(\check{J}, \check{X})$ is a solution;
- on the other hand, it is strictly less interesting than $(J, X)$ — because it is a restriction.

**Definition.** Let $(J_1, X_1)$, $(J_2, X_2)$ be two solutions to an equation (possibly with initial condition). One says that $(J_2, X_2)$ extends $(J_1, X_1)$ if:

- $J_1 \subseteq J_2$;
- $X_1 = X_{2|J_1}$, i.e. $\forall t \in J_1,\ X_1(t) = X_2(t)$.

A maximal solution is one which does not have a proper extension.

**Remark.** A global solution is always maximal, but the converse can fail as we shall see.

The following derives from set-theoretic abstract non-sense.

**Fact.** Every solution extends to a maximal one.

As a consequence we now focus on maximal solutions; bear in mind that they need not be global.

### II.1.2. Phenomena

In the linear world all is fine:

- a Cauchy problem has a unique solution;
- all (maximal) solutions are global.

Neither is true in the non-linear world.

- **A Cauchy problem without uniqueness**

**Example.** Consider the Cauchy problem:

$$x'(t) = 2\sqrt{|x(t)|} \qquad \text{on } I = \mathbb{R}$$

with initial condition $x(0) = 0$. We shall construct *infinitely many* solutions.
For $T \in \mathbb{R}_{\geqslant 0}$ let:

$$x_T(t) = \begin{cases} 0 & \text{if} \quad t \leqslant T \\ (t-T)^2 & \text{if} \quad t \geqslant T \end{cases}$$

We make the following observations:

- $x_T$ is defined on $\mathbb{R}$;
- it is continuous on $(-\infty, T)$ and $(T, +\infty)$; it also is continuous at $T$, so it is on $(-\infty, +\infty)$;
- it is actually differentiable on $(-\infty, T)$ and $(T, +\infty)$; at $T$ it admits a left semi-derivative:
$$(x_T)'_- = \lim_{h \to 0^-} 0 = 0$$

and a right semi-derivative:

$$(x_T)-_+ = \lim_{h\to 0^+} \frac{h^2}{h} = 0$$

so it is actually differentiable on $\mathbb{R}$;

- for all $t \in \mathbb{R}$ one has $x'_T(t) = 2\sqrt{|x(t)|}$;
- $x_T(0) = 0$.

So any $x_T$ for $T \geqslant 0$ is a solution: the Cauchy problem has infinitely many.

Under some special assumptions on $G$, the Cauchy-Lipschitz Theorem will rule out this situation.

- **A differential equation with no global solution**

**Example.** Consider the equation:

$$y'(t) = \frac{1}{2y(t)} \qquad \text{on } I = \mathbb{R}$$

We shall prove that no solution is defined on $\mathbb{R}$.

Before we start: of course we think about $\sqrt{t}$. But notice that this function is not defined on $I = \mathbb{R}$.

So let $(J, y)$ be a maximal solution to the equation; we prove $J \subsetneq \mathbb{R}$. First, since $y$ is continuous and does not vanish, it keeps a constant sign; we may assume that $y$ remains positive on $J$ (the other case is similar).

Observe how $2y'(t)y(t) = 1$, so $(y^2(t))' = 1$; fixing $t_1 \in J$ one finds:

$$\forall t \in J, \quad y^2(t) = (t - t_1) + y^2(t_1)$$

This forces $t - t_1 + y^2(t_1) \geqslant 0$, hence $t \geqslant t_1 - y^2(t_1)$. In particular, $J \subseteq (t_1 - y^2(t_1), +\infty) \subsetneq \mathbb{R}$.

There will be no remedy to this. "Explosion phenomena" are unavoidable; they however give rise to a decent theory.

### II.1.3. Euler's method

Euler's method is an approximation method based on discretisation, i.e. turning the continuous problem of solving a differential equation into the *discrete* problem of solving a difference equation.

On some subinterval $[a, b] \subseteq I$ proceed as follows:

- choose a "step $h$" (the smaller it is, the longer your computations will be, but the better you hope your approximation to be);
- divide the interval $[a, b]$ into $N = \frac{b-a}{h}$ equal subinterval with endpoints:

$$a = t_0 < t_1 = t_0 + h < \cdots < t_N = t_0 + Nh = b.$$

- Now starting from the initial condition $X_0$, iteratively compute the values:

$$X_{i+1} = X_i + hG(t_i, X_i)$$

Reason: Euler's *hope* is that if $h$ is small enough, then $t_{i+1} - t_i = h$ should be small enough to write:

$$G(t_i, X(t_i)) = X'(t_i) = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon}(X(t_i + \varepsilon) - X(t_i)) \approx \frac{1}{h}(X(t_{i+1}) - X(t_i))$$

which results in the above construction.

- Then construct a piecewise affine map $Y_h$ such that $Y_h(t_i) = X_i$ for all $i = 0 \ldots N$. One may actually write down the expression:

$$\text{for } t \in [t_i, t_{i+1}], \text{ let } Y_h(t) = \frac{t - t_i}{t_{i+1} - t_i}(X_{i+1} - X_i) + X_i$$

- Hope: as $h \to 0$, functions $Y_h$ are desired to converge to a function, which in turn is desired to be a solution.

Sadly enough, this is wishful thinking, and not true in general: the method need not converge without strong assumptions, of Lipschitz type.

## II.2. Existence: the Cauchy-Peano theorem

**Theorem** (Peano)**.** Consider a Cauchy problem *where $G$ is continuous.* Then there exists a solution $(J, X)$.

**Remarks.**

- As we know from the example of $x'(t) = 2\sqrt{|x(t)|}$ (check that the theorem applies to it!), uniqueness cannot be expected.
- As we know from the example of $y'(t) = \frac{1}{2y(t)}$ (check that the theorem applies to it!), globality should not be expected.
- Let us push the limits of the result: an implicit differential equation is one of the form $H(t, X(t), X'(t)) = 0$ (our original definition therefore focuses on explicit equations).

  Here is an example with continuous $H$ and yet no solution at all. Let $H(t, x_1, x_2) = x_1 x_2 + t$, i.e. consider equation:

$$x(t)x'(t) + t = 0$$

  with initial condition $x(0) = 0$. Suppose that $(J, x)$ is a solution. Observe how $x^2(t) + t^2$ is a constant on $J$; it must be 0, which forces $t = 0$ and the open interval $J$ is a singleton, a contradiction.

- Peano's theorem is a beautiful existence result based on Euler's method; in this sense it can arguably be called constructive. However, since there is no way to effectively predict which subsequence of approximations will converge, it is useless in practice.

**Proof**. The proof will take us to functional analysis. Remember that if $[a, b] \subseteq \mathbb{R}$ is a compact interval, then any continuous function from $[a, b]$ to the reals is bounded, and reaches its supremum. Moreover, any continuous function from $[a, b]$ to a metric space is uniformly continuous.

As a consequence of the first, the vector space:

$$C^0\left([a, b], \mathbb{R}^d\right) = \left\{f : [a, b] \to \mathbb{R}^d \text{ a continuous map}\right\}$$

can be equipped with the uniform norm

$$\|f\|_\infty = \sup_{t \in [a,b]} \|f(t)\| = \max_{[a,b]} \|f\|$$

Here is a fact from functional analysis we admit.

**Theorem** (Ascoli)**.** Let $[a, b] \subseteq \mathbb{R}$ be a compact interval, $E = C^0([a, b], \mathbb{R}^d)$ with norm $\| \cdot \|_\infty$, and $(f_n)$ be a sequence of elements of $E$. Suppose that:

- $(f_n)$ is equi-bounded, i.e.:

$$\exists A \in \mathbb{R}, \quad \forall n \in \mathbb{N}, \quad \|f\|_\infty \leqslant A;$$

- $(f_n)$ is equi-uniformly continuous, i.e.:

$$\begin{aligned} &\forall \varepsilon > 0 \exists \eta > 0, \forall n \in \mathbb{N}, \forall (s_1, s_2) \in [a, b]^2 \\ &\quad |s_1 - s_2| < \eta \Rightarrow \|f_n(s_1) - f_n(s_2)\| < \varepsilon. \end{aligned}$$

Then $(f_n)$ has a converging subsequence (convergence being of course wrt $\| \cdot \|_\infty$, i.e. uniform convergence).

**Remark.** One would love to say that $(f_n)$ is *uniformly bounded*, which certainly has a clear meaning; but for consistency one should then say that $(f_n)$ is *uniformly uniformly continuous*, which sounds strange.

The first step involves restricting ourselves to a subset of an important form — so important that there will be a general definition.

**Step 1.** There exists a security cylinder.

**Definition.** A security cylinder for a Cauchy problem is a compact set $K = [t_0 - \alpha, t_0 + \alpha] \times \overline{B}(X_0, r) \subseteq I \times U$ such one has $\alpha M < r$, where:

$$M = \sup_{(s,Y) \in K} \|G(s, Y)\|$$

The idea behind a security cylinder is that if $(J, X)$ is a solution with initial condition, then $X(t)$ remains inside $\overline{B}(X_0, r)$ while $t \in J \cap [t_0 - \alpha, t_0 + \alpha]$.

Intuitively, while $|t - t_0| < \alpha$, one has $\|X'(t)\| = \|G(t, X(t))\| \leqslant M$, so that $\|X(t) - X(t_0)\| \leqslant M|t - t_0| \leqslant M_\alpha < r$. (An exercise is to turn this into a rigorous proof.)

**Proof of Step 1**. The open set $I \times U$ contains $(t_0, X_0)$, so it contains a small compact neighborhood of the form $K = [t_0 - \alpha, t_0 + \alpha] \times \overline{B}(X_0, r)$. Since $G$ is continuous on $M$ (actually everywhere), it is bounded in norm by say $M$. We can always suppose that $\alpha < \frac{r}{M}$. $\lozenge$

**Step 2.** For all $\varepsilon > 0$, there is a nice $\varepsilon$-approximate solution on $[t_0 - \alpha, t_0 + \alpha]$.

Here again the step requires a formal definition.

**<u>Definition</u>.** An <u>$\varepsilon$-approximate solution</u> is a continuous map $Y_\varepsilon : [t_0 - \alpha, t_0 + \alpha] \to \mathbb{R}^d$ which is *right*-differentiable and satisfies:

$$\forall t \in [t_0 - \alpha, t_0 + \alpha), \quad \left\|(Y_\varepsilon)'_+(t) - G(t, Y_\varepsilon(t))\right\| < \varepsilon$$

Typically Euler's method gives $\varepsilon$-approximate solutions. Notice that semi-differentiability is unavoidable when one works with piecewise affine functions.

**Proof of Step 2**. Fix $\varepsilon > 0$. Since $G$ is continuous on the compact $K$, it is uniformly continuous there; so there exists $\eta$ such that:

$$\begin{aligned}&\forall (s_1, X_1), (s_2, X_2) \in K,\\ &\qquad |s_1 - s_2| \wedge \|X_1 - X_2\| < \eta \Rightarrow \|G(s_1, X_1) - G(s_2, X_2)\| < \varepsilon;\end{aligned}$$

notice that $\eta$ depends on $\varepsilon$.

Let us apply Euler's method. We divide the interval $[t_0, t_0 + \alpha]$ regularly with step $h = \min(\eta, \frac{\eta}{M})$; notice that $h$ depends on $\varepsilon$. This results in $N = \frac{\alpha}{h}$ subintervals with endpoints:

$$t_0 < t_1 < \cdots < t_N = t_0 + \alpha;$$

here again, $N$ and the various $t_i$'s depend on $\varepsilon$ (if one were after perfect clarity one would write $t_{\varepsilon,i}$).

As in Euler's method, we define inductively:

$$X_{i+1} = X_i + hG(t_i, X_i)$$

since $X_0$ is given by the initial condition (here again, one could write $X_{\varepsilon,i}$ for full clarity).

Observe that by induction:

$$\forall i = 0 \dots N, \quad \|X_i - X_0\| \leqslant ihM \leqslant NhM = \alpha M < r$$

The last two result from $Nh = \alpha$ as in any subdivision, and the requirement that $\alpha M < r$ is in the definition of a security cylinder. The first inequality is induction properly speaking, since if the claim holds at $i$, then $(t_i, X_i) \in K$ so $\|G(t_i, X_i)\| \leqslant M$ and:

$$\|X_{i+1} - X_0\| \leqslant \|X_{i+1} - X_i\| + \|X_i - X_0\| \leqslant hM + ihM = (i+1)hM$$

This means that *each* $(t_i, X_i)$ *is in the security cylinder.*

As in Euler's method, we define $Y_\varepsilon$ as a piecewise affine function with $Y_\varepsilon(t_i) = X_i$, namely:

$$\text{for } t \in [t_i, t_{i+1}), \quad Y_\varepsilon(t) = \frac{t - t_i}{t_{i+1} - t_i}(X_{i+1} - X_i) + X_i;$$

in particular, while $t \in [t_0, t_0 + \alpha]$, $(t, Y_\varepsilon(t))$ *remains in the security cylinder.*

As any continuous piecewise affine function, $Y_\varepsilon$ is left- and right-differentiable everywhere; here, for $t \in [t_i, t_{i+1})$, one has:

$$(Y_\varepsilon)'_+(t) = \frac{1}{t_{i+1} - t_i}(X_{i+1} - X_i) = \frac{1}{h} hG(t_i, X_i) = G(t_i, X_i)$$

so that, still for $t \in [t_i, t_{i+1})$:

$$\left\|(Y_\varepsilon)'_+(t) - G(t, Y_\varepsilon(t))\right\| = \|G(t_i, X_i) - G(t, Y_\varepsilon(t))\|$$

Now on the one hand, $|t - t_i| < h \leqslant \eta$, and on the other hand $\|X_i - Y_\varepsilon(t)\| \leqslant \|X_i - X_{i+1}\| \leqslant hM \leqslant \eta$, so by definition of $\eta$ one has:

$$\left\|(Y_\varepsilon)'_+(t) - G(t, Y_\varepsilon(t))\right\| < \varepsilon$$

This defines an $\varepsilon$-approximate solution on $[t_0, t_0 + \alpha]$. Work on $[t_0 - \alpha, t_0)$ using a similar procedure. ◊

As we know there is in general no guarantee that Euler's method will result in anything sensible — nor even in anything at all. The key idea to complete the proof of Peano's theorem is to use Ascoli's result from functional analysis to force convergence of a subsequence.

**Step 3.** There exists a solution.

**Proof**. The idea will be to let $\varepsilon = \frac{1}{n}$ and let $n$ go to $\infty$; the problem of course is to find a converging subsequence. Applying Ascoli's theorem requires equi-boundedness and equi-uniform continuity, so we must return to the function $Y_\varepsilon$ we constructed and see that some of its properties do not depend on $\varepsilon$.

Let $t \in [t_0, t_0 + \alpha)$, then there is $i$ with $t \in [t_i, t_{i+1})$, so:

$$\left\|(Y_\varepsilon)'_+(t)\right\| = \|G(t_i, X_i)\| \leqslant M,$$

with a similar property for any $t \in (t_0 - \alpha, t_0]$. This means that $Y_\varepsilon$ is $M$-Lipschitz. As this does not depend on $\varepsilon$, the family $\{Y_\varepsilon : \varepsilon > 0\}$ is equi-uniformly continuous. Moreover, integrating between $t_0$ and $t$:

$$\|Y_\varepsilon(t) - Y_\varepsilon(t_0)\| \leqslant |t - t_0|M \leqslant \alpha M < r,$$

so $\|Y_\varepsilon(t)\| \leqslant \|X_0\| + r$ and $\|Y_\varepsilon\|_\infty \leqslant \|X_0\| + M$, where the supremum was taken over $[t_0 - \alpha, t_0 + \alpha]$. So the family $\{Y_\varepsilon : \varepsilon > 0\}$ is equi-bounded.

For $n \in \mathbb{N}\backslash\{0\}$ we let $Z_n = Y_{\frac{1}{n}}$; each $Z_n$ is a $\frac{1}{n}$-approximate solution, and the sequence $(Z_n)$ satisfies the assumptions of Ascoli's theorem. There exists therefore a converging subsequence $(Z_{\varphi(n)})$ with limit say $Z$.

We claim that $Z$ is a solution; this involves one final idea. Notice that in general, $(J, X)$ is a solution to a Cauchy problem iff on $J$ one has:

$$X(t) = X_0 + \int_{t_0}^{t} G(s, X(s))ds$$

We shall prove that $Z$ satisfies this functional relation.

Consider the error term:

$$R_n(t) = Z_n(t) - X_0 - \int_{t_0}^{t} G(s, Z_n(s))ds$$

Observe how $R_n(t_0) = Z_n(t_0) - X_0 = 0$. Moreover $R_n$ is right-differentiable everywhere; letting $\varepsilon = \frac{1}{n}$ so that $Z_n = Y_\varepsilon$ and returning to earlier computations:

$$\begin{aligned}
\left\|(R_n)'_+(t)\right\| &= \left\|(Z_n)'_+(t) - G(t, Z_n(t))\right\| \\
&= \left\|(Y_\varepsilon)'_+(t) - G(t, Y_\varepsilon(t))\right\| \\
&\leqslant \varepsilon = \frac{1}{n}
\end{aligned}$$

Integrating, this gives $\|R_n(t)\| = \|R_n(t) - R_n(t_0)\| \leqslant \frac{1}{n}\alpha$, so $\|R_n\|_\infty \leqslant \frac{1}{n}\alpha$, and $R_n \to 0$, even uniformly so.

Now since $Z_{\varphi(n)} \to Z$ uniformly and since $G$ is uniformly continuous,

$$G(t, Z_{\varphi(n)}(t)) \to G(t, Z(t))$$

uniformly; hence integrating on the compact $[t_0, t]$:

$$\int_{t_0}^{t} G(s, Z_{\varphi(n)}(s))ds \to \int_{t_0}^{t} G(s, Z(s))ds$$

Here convergence is only simple (i.e. pointwise). Another way to obtain it is by measure theory and Lebesgue's dominated convergence theorem; the latter option does not use uniform continuity of $G$, but of course the possibility to find a dominating function; it is an exercise to check the assumptions.

In any case, one has:

$$R_{\varphi(n)}(t) \to Z(t) - X_0 - \int_{t_0}^{t} G(s, Z(s))ds$$

where convergence is simple/pointwise. But we already know the limit is 0:

so for all $t \in [t_0 - \alpha, t_0 + \alpha]$, one has:

$$Z(t) - X_0 - \int_{t_0}^{t} G(s, Z(s))ds = 0,$$

proving that $Z$ is a solution to the Cauchy problem. $\Diamond$

This completes the proof of Peano's Theorem. $\square$

**Remark.** The practical value of Peano's theorem is zero. As one sees from the proof, if one runs Euler's method with steps going to 0, getting always better approximate solutions $Z_n$, all we know is that there exists a subsequence which converges uniformly to a solution.

But there are continuum-many such subsequences: it is out of question to try them all.

# II.3. Existence *and* uniqueness

## II.3.1. The result (local version)

As oppposed to Peano's result, the Cauchy-Lipschitz theorem, also known as the Picard-Lindelöf theorem, is fundamental both in theory and in practice; it is so important that we shall prove it *twice*.

**Theorem** (Cauchy-Lipschitz/Picard-Lindelöf: local version)**.** Consider a Cauchy problem:

$$\begin{cases} X'(t) = G(t, X(t)) \text{ for } t \in I \\ X(t_0) = X_0 \end{cases}$$

Suppose as always that $G$ is continuous.

Moreover suppose that there is a security cylinder $K = \overline{J} \times \overline{B}(X_0, r) \subseteq I \times U$ of $(t_0, X_0)$ on which $G$ is Lipschitz with respect to $X$, i.e.:

$$\exists k > 0 \ \forall t \in J \ \forall (X_1, X_2) \in B(X_0, r), \quad \|G(t, X_1) - G(t, X_2)\| \leqslant k \|X_1 - X_2\|$$

Then the problem has a unique (local) solution on this neighborhood.

**Remarks.**

- One sometimes says that $G$ is locally Lipschitz in the second variable.
- Here is a trivial criterion: if $G$ is $C^1$ (has continuous partial derivatives), then it is locally Lipschitz.

  Hence to apply the theorem it suffices to check that $G(t, V)$ is $C^0$ as a function of $(t, V)$ and $C^1$ as a function of $V$.
- Return to example $x' = 2\sqrt{|x(t)|}$ on $\mathbb{R}$. The function $\sqrt{|.|}$, though continuous, is *not* Lipschitz on any neighborhood of 0 (the derivative

goes to $+\infty$). So lack of uniqueness in this case is not against the Cauchy-Lipschitz theorem.

- Still, the Cauchy-Lipschitz theorem does *not* guarantee the existence of global solutions. Return to example $y' = \frac{1}{2y(t)}$ on $\mathbb{R}$. The function $\frac{1}{2\cdot}$ is locally Lipschitz in its variable, so the theorem applies. As we know, no solution is global though.

  As a corollary we shall give a special version where solutions are global.

## II.3.2. Grönwall's Lemma

To prove the Cauchy-Lipschitz theorem we shall return to Euler's method. In the proof of Peano's theorem it was unclear whether the whole approximating sequence would converge to a solution, nor if the convergence speed would be satisfactory as we had no explicit control on the error term. Under a Lipschitz assumption this is possible. We shall use a fundamental result.

**Lemma** (Grönwall)**.** Let $f, g : I \to \mathbb{R}_{\geqslant 0}$ be continuous and $c \in \mathbb{R}$ be such that:

$$\forall t \geqslant t_0, \quad f(t) \leqslant c + \int_{t_0}^{t} f(s)g(s)ds$$

Then:

$$\forall t \geqslant t_0, \quad f(t) \leqslant c \exp\left(\int_{t_0}^{t} g(s)ds\right)$$

**Proof.** The proof is pure magic. Let:

$$h(t) = \frac{c + \int_{t_0}^{t} f(s)g(s)ds}{\exp\left(\int_{t_0}^{t} g(s)ds\right)}$$

This function is continuous at $t_0$, differentiable on $I$, and for $t \geqslant t_0$ the derivative is:

$$h'(t) = \frac{f(t)g(t) - \left(c + \int_{t_0}^{t} f(s)g(s)ds\right) \cdot g(t)}{\exp\left(\int_{t_0}^{t} g(s)ds\right)} \leqslant 0$$

So, always for $t \geqslant t_0$, one has $h(t) \leqslant h(t_0)$. Therefore when $t \geqslant t_0$:

$$f(t) \leqslant c + \int_{t_0}^{t} f(s)g(s)ds = h(t) \cdot \exp\left(\int_{t_0}^{t} g(s)ds\right) \leqslant h(t_0) \cdot \exp\left(\int_{t_0}^{t} g(s)ds\right);$$

as $h(t_0) = c$ we are done. □

## II.3.3. Proof of the Cauchy-Lipschitz theorem

Let us apply this striking result to Euler's method, in the Lipschitz case.

**First proof of the Cauchy-Lipschitz theorem**.

**Step 1.** Light refreshments.

**Proof**. Let us refresh our memories:

- we are given a Cauchy problem,
- and a security cylinder $K = \overline{J} \times \overline{B}(X_0, r)$, i.e. $J = (t_0 - \alpha, t_0 + \alpha)$ and $M = \sup_K \|G\|$ are subject to $\alpha M < r$.
- The assumption is that $G$ is locally Lipschitz in $X$ on the cylinder: there is $k$ such that for all $t \in \overline{J}$, the map $G(t, \cdot)$ is $k$-Lipschitz on $\overline{B}(X_0, r)$.

We shall run Euler's method and prove that it converges. Let us again refresh our memories:

- for $\varepsilon > 0$, we let $n = \frac{1}{\varepsilon}$, and then divide $[t_0, t_0 + \alpha]$ into $n$ subintervals with step $h = \frac{\alpha}{n}$ and endpoints $t_0 < t_1 < \cdots < t_n = t_0 + \alpha$; be careful that in full notation this should be $n_\varepsilon$, $h_\varepsilon$ and $t_{\varepsilon,0}, \ldots t_{\varepsilon,n_\varepsilon}$ but this would become unreadable;
- we then compute $X_{i+1} = X_i + hG(t_i, X_i)$ inductively; here again it should be the family $X_{\varepsilon,0}, \ldots X_{\varepsilon,n_\varepsilon}$;
- finally we define $Y_\varepsilon$ to be the continuous, piecewise affine function taking $Y_\varepsilon(t_i) = X_i$;
- as we know, $Y_\varepsilon$ is an $\varepsilon$-approximate solution: it is right-differentiable and
$$\left\|(Y_\varepsilon)'_+(t) - G(t, Y_\varepsilon(t))\right\| < \varepsilon \qquad \lozenge$$

Peano's theorem used Ascoli's result from functional analysis to force one subsequence of $(Z_n) = (Y_{\frac{1}{n}})$ to converge to a solution. Here the situation is undoubtedly better thanks to the Lipschitz assumption (not used so far).

**Step 2.** If $Y_1, Y_2$ are $\varepsilon_1$-, resp. $\varepsilon_2$-approximate solutions (with exact initial condition), then on their common definition interval:

$$\|Y_1(t) - Y_2(t)\| \leqslant \alpha(\varepsilon_1 + \varepsilon_2) \exp(k(t - t_0)).$$

**Proof**. Let $Z(t) = Y_1(t) - Y_2(t)$, a right-differentiable map. Write:

$$\begin{aligned} Z'_+(t) &= (Y_1)'_+(t) - (Y_2)'_+(t) \\ &= \left[(Y_1)'_+(t) - G(t, Y_1(t))\right] - \left[(Y_2)'_+(t) - G(t, Y_2(t))\right] \\ &\quad + \left[G(t, Y_1(t)) - G(t, Y_2(t))\right], \end{aligned}$$

so that using the $k$-Lipschitz property:

$$\|Z'_+(t)\| \leqslant \varepsilon_1 + \varepsilon_2 + k\|Y_1(t) - Y_2(t)\| = (\varepsilon_1 + \varepsilon_2) + k\|Z(t)\|$$

Integrating and bearing in mind $Z(t_0) = Y_1(t_0) - Y_2(t_0) = X_0 - X_0 = 0$, this yields:

$$\|Z(t)\| \leqslant \int_{t_0}^{t} (\varepsilon_1 + \varepsilon_2) + k\|Z(s)\|ds \leqslant \alpha(\varepsilon_1 + \varepsilon_2) + k\int_{t_0}^{t} \|Z(s)\|ds$$

Introduce maps $f(t) = \|Z(t)\|$ and $g(t) = k$. By Grönwall's lemma,

$$\|Z(t)\| \leqslant \alpha(\varepsilon_1 + \varepsilon_2)\exp(k(t - t_0)),$$

the desired estimate. $\Diamond$

(As a side remark: one can also write a variant where the initial condition is only $\varepsilon_i$-satisfied, i.e. $\|Y_i(t_0) - X_0\| < \varepsilon_i$.)

The above estimate proves both existence and uniqueness.

- Uniqueness first.

  If $Y_1, Y_2$ are two exact solutions, then all the above remains valid with $\varepsilon_1 = \varepsilon_2 = 0$, so $Z$ is the null function, meaning $Y_1 = Y_2$.

- Now to existence.

  Take any sequence $\varepsilon_n \to 0$ (for instance $\varepsilon_n = \frac{1}{n}$); for simplicity let $Y_n = Y_{\varepsilon_n}$ and $Z_{n,m} = Y_n - Y_m$. By our last estimate, $\|Z_{n,m}(t)\| \to 0$ as $n, m \to \infty$; as a matter of fact, if $t$ remains in a compact interval, convergence is uniform. By completeness of the relevant function space, it means that $(Y_n)$ converges (and uniformly on compact intervals) to a function $Y$.

  Rewriting the differential equation with initial condition as the integral equation $X(t) = X_0 + \int_{t_0}^{t} G(s, X(s))ds$, we see that $Y$ is a solution. $\square$

**Remarks.**

- As opposed to Peano's theorem, the whole Euler method will converge.
- The proof works only because thanks to the Lipschitz property error terms can be bounded.

# II.4. More on Cauchy-Lipschitz

## II.4.1. Global version

As a matter of fact the proof gives a bit more.

**Theorem** (Cauchy-Lipschitz: global version)**.** Consider a Cauchy problem with continuous $G$.

Suppose that there is a function $k : I \to \mathbb{R}_{\geqslant 0}$ such that for all $t \in I$, the map $G(t, \cdot)$ is globally $k(t)$-Lipschitz, i.e.:

$$\forall t \in I, \forall (X_1, X_2) \in U^2, \quad \|G(t, X_1) - G(t, X_2)\| \leqslant k(t) \cdot \|X_1 - X_2\|.$$

Then the Cauchy problem has a unique maximal solution *which is global.*

**Proof.** In the notation of the proof above, let $g(s) = k(s)$. We then find:

$$\|Z(t)\| \leqslant \int_{t_0}^{t} (\varepsilon_1 + \varepsilon_2) + \|Z(s)\| g(s) ds \leqslant |t - t_0|(\varepsilon_1 + \varepsilon_2) + \int_{t_0}^{t} \|Z(s)\| g(s) ds,$$

so by Grönwall's lemma

$$\|Z(t)\| \leqslant |t - t_0|(\varepsilon_1 + \varepsilon_2) \cdot \exp\left(\int_{t_0}^{t} g(s) ds\right).$$

In the case of $Z(t) = Y_n(t) - Z_m(t)$, while $t$ remains at *bounded* distance from $t_0$, we find convergence again to a solution. $\square$

**Examples.** Be extremely careful with the theorem: to have a global solution, one needs for all $t$ function $G(t, \cdot)$ to be *globally* Lipschitz.

- Consider equation $x'(t) = x(t)$ with initial condition $x(0) = 1$. As we know from explicit resolution, there is a unique maximal solution, which is global. Indeed, here one has $G(t, x) = x$, which at every $t$ is 1-Lipschitz.

- Now consider equation $x'(t) = x^2(t)$ with $x(0) = 1$. Here $G(t, x) = x^2$; for any $t$, around every $x_1$, it is locally $(2|x_1| + 1)$–Lipschitz (we add a little something to make sure it is true), but this depends on $x_1$; as a matter of fact $G(t, \cdot)$ is *not* globally Lipschitz.

  This is the reason why there is *no global solution.* One can be easily convinced: on the one hand $x(t) = \frac{1}{(1-x)}$ is a maximal solution, on the other hand it is not global. By local uniqueness (a consequence of the local Cauchy-Lipschitz theorem), or the "disjunction property" below, there is no global solution.

- The global version holds in particular for *linear* differential equations, since in that case:

  $$\|G(t, X_1) - G(t, X_2)\| = \|A(t) \cdot (X_1 - X_2)\| \leqslant |||A(t)||| \cdot \|X_1 - X_2\|$$

  so taking $k(t) = |||A(t)|||$ is suitable.

## II.4.2. The disjunction property

Here is an important corollary to the Cauchy-Lipschitz theorem, saying that two distinct maximal solutions can never intersect.

**Corollary.** Suppose $X'(t) = G(t, X(t))$ is a differential equation on $I$ such that for every $t$, $G(t, \cdot)$ is $k(t)$-Lipschitz for some number $k(t)$.

Let $(J_1, X_1)$ and $(J_2, X_2)$ be two maximal solutions. If there is $t_0 \in J_1 \cap J_2$ such that $X_1(t_0) = X_2(t_0)$, then $J_1 = J_2$ and $X_1 = X_2$.

**Proof.**

**Lemma** (common extension criterion)**.** Let $(J_1, X_1)$ and $(J_2, X_2)$ be two solutions of any differential equation. Suppose that $\check{J} = J_1 \cap J_2$ is non-empty, and $X_{1|\check{J}} = X_{2|\check{J}}$, i.e. for $t \in J_1 \cap J_2$ one has $X_1(t) = X_2(t)$.

Then $(J_1, X_1)$ and $(J_2, X_2)$ have a common extension.

**Proof.** Let $\hat{J} = J_1 \cup J_2$, which is:

- non-empty, as neither $J_1$ nor $J_2$ can be;
- open, as both $J_1$ and $J_2$ are;
- connected, since it is the union of two connected sets *with non-trivial intersection.*

Hence $\hat{J}$ is a non-empty open subinterval of $\mathbb{R}$. On $\hat{J}$ we define a function $\hat{X}$ by:

$$\hat{X}(t) = \begin{cases} X_1(t) & \text{if} \quad t \in J_1 \\ X_2(t) & \text{if} \quad t \in J_2 \end{cases};$$

this is well-defined since by assumption, $X_1$ and $X_2$ agree on $J_1 \cap J_2$.

It is then obvious that $\hat{X}$ is differentiable on $\hat{J}$, and a solution of the equation there. By construction, it extends both $(J_1, X_1)$ and $(J_2, X_2)$. $\Diamond$

If in the above $(J_1, X_1)$ and $(J_2, X_2)$ are supposed to be maximal, then by definition of a maximal solution one must have $J_1 = \hat{J}$ and $X_1 = \hat{X}$; but also $J_2 = \hat{J}$ and $X_2 = \hat{X}$, which will prove the conclusion.

Hence to prove the corollary, it suffices to show:

> Same assumptions as in the corollary, but $X_1$ and $X_2$ are *not* supposed to be maximal. Then $X_1$ and $X_2$ agree on $J_1 \cap J_2$.

**Proof of this claim.** We let $\check{J} = J_1 \cap J_2$, which is:

- non-empty as $t_0 \in J_1 \cap J_2$;
- open as both $J_1$ and $J_2$ are;
- connected since it is the intersection of two convex sets, hence a convex set, and convex subsets are always connected.

So $\check{J}$ is a non-empty, open subinterval. Also introduce:

$$A = \left\{ t \in \check{J} : X_1(t) = X_2(t) \right\},$$

which is non-empty as $t_0 \in A$ by assumption. We must prove that $A = \check{J}$. We argue using connectedness of $\check{J}$.

- $A$ is closed in $\check{J}$: because $A = Y^{-1}(\{0\})$, where $Y : \check{J} \to \mathbb{R}^d$ takes $t$ to $X_1(t) - X_2(t)$. Since $Y$ is continuous and $\{0\}$ is closed in $\mathbb{R}^d$, $A$ is closed in $\check{J}$.
- $A$ is open in $\check{J}$: let $t_a \in A$ and $X_a = X_1(t_a) = X_2(t_a)$ (since $t_a \in A$). We may apply the Cauchy-Lipschitz theorem with initial condition $(t_a, X_a)$ and find that on some $J_a = (t_a - \alpha, t_a + \alpha)$, there is a unique solution with $X(t_a) = X_a$.

  But both the restrictions $X_{1|J_a}$ and $X_{2|J_a}$ meet the requirements: so they are equal. As a conclusion, $X_1$ and $X_2$ agree on $J_a$, meaning $J_a \subseteq A$: hence $A$ is an open subset.

**Remark.** We even proved that $A$ is open in $\mathbb{R}$; it is not necessarily true that $A$ is closed in $\mathbb{R}$ (but it is in $\check{J}$, which suffices to conclude as follows).

Since $\check{J}$ is connected, its only clopen (=closed and open) subsets are $\varnothing$ and $\check{J}$ itself. But $A \neq \varnothing$, so $A = \check{J}$, as desired. $\lozenge$

This concludes the proof of the Corollary. $\square$

### II.4.3. Alternate proof of Cauchy-Lipschitz

**Second proof of the Cauchy-Lipschitz theorem**. Existence could be derived from Peano's theorem, so only uniqueness would remain; we shall however give a completely different argument.

Let us return to an earlier idea: $(J, X)$ is a solution iff on $J$ one has:

$$X(t) = X_0 + \int_{t_0}^{t} G(s, X(s))ds$$

This means that $X$ is a fixed-point of some operator from functional analysis:

$$\begin{array}{rccc} T: & C^0\left(\overline{J}, \overline{B}(X_0, r)\right) & \to & C^0\left(\overline{J}, \mathbb{R}^d\right) \\ & Y & \mapsto & X_0 + \int_{t_0}^{t} G(s, Y(s))ds \end{array}$$

**Step 1.** The closed set $C = C^0\left(\overline{J}, \overline{B}(X_0, r)\right)$ is stable under $T$.

**Proof**. As in the definition of a security cylinder, we let $\overline{J} = [t_0 - \alpha, t_0 + \alpha]$

and:

$$M = \sup_K \|G\|$$

with the requirement $\alpha M < r$.

We prove stability. If $Y \in C$, i.e. if $Y$ is a continuous function on $\overline{J}$ taking values in $\overline{B}(X_0, r)$, then for all $t \in \overline{J}$:

$$\|(T(Y))(t) - X_0\| = \left\| \int_{t_0}^{t} G(s, Y(s))ds \right\| \leqslant \int_{t_0}^{t} \|G(s, Y(s))\| \, ds$$

But since we remain in the security cylinder, $\|G(s, Y(s))\| \leqslant M$ at all times, and therefore:

$$\|(T(Y))(t) - X_0\| \leqslant |t - t_0| M \leqslant \alpha M < r$$

so $T(Y)$, a continuous map on $\overline{J}$, takes values in $B(X_0, r)$: hence $T(Y) \in C$.

This step did not use the local Lipschitz condition. ◊

**Step 2.** $T : C \to C$ is a continuous map and has a power which is a contraction mapping.

**Definition.** Let $(C, d)$ be a metric space. A map $S : C \to C$ is called a contraction mapping if there is $0 \leqslant \ell < 1$ such that:

$$\forall (c_1, c_2) \in C^2, \quad d(S(c_1), S(c_2)) \leqslant \ell d(c_1, c_2)$$

**Proof of Step 2**. Let $Y_1, Y_2 \in C$. Again this means that we have two continuous maps on $\overline{J}$ taking values "close to $X_0$". Then for any $t \in \overline{J}$ one has:

$$\begin{aligned} \|T(Y_1)(t) - T(Y_2)(t)\| &= \left\| \int_{t_0}^{t} G(s, Y_1(s)) - G(s, Y_2(s))ds \right\| \\ &\leqslant \int_{t_0}^{t} \|G(s, Y_1(s)) - G(s, Y_2(s))\| \, ds \end{aligned}$$

Now we use the local Lipschitz character: since we remain in the security cylinder, for all $s \in [t_0, t]$ one has $\|G(s, Y_1(s)) - G(s, Y_2(s))\| \leqslant k \|Y_1(s) - Y_2(s)\| \leqslant k \|Y_1 - Y_2\| \infty$, which implies:

$$\|T(Y_1)(t) - T(Y_2)(t)\| \leqslant |t - t_0| k \|Y_1 - Y_2\|_\infty$$

We can already take the supremum for $t \in \overline{J}$ and find $\|T(Y_1) - T(Y_2)\|_\infty \leqslant \alpha k \|Y_1 - Y_2\|_\infty$: hence $T$ is $\alpha k$-Lipschitz. This guarantees continuity, but one could well have $\alpha k \geqslant 1$, which is not satisfactory for a contraction mapping.

A possibility could be to reduce $\alpha$, but then we would not prove our exact statement (as we would be working on a sub-cylinder of the one given to us). So let us refine the computation:

$$\|T(Y_1)(t) - T(Y_2)(t)\| \leqslant |t - t_0| k \|Y_1 - Y_2\|_\infty,$$

so

$$\begin{aligned}\left\|T^2(Y_1)(t) - T^2(Y_2)(t)\right\| &\leqslant \int_{t_0}^{t} \|G(s, T(Y_1)(s)) - G(s, T(Y_2)(s))\| \, ds \\ &\leqslant \int_{t_0}^{t} k^2 |s - t_0| \|Y_1 - Y_2\|_\infty ds \\ &= \frac{1}{2} k^2 |t - t_0|^2 \|Y_1 - Y_2\|_\infty .\end{aligned}$$

One can then prove by induction:

$$\|T^n(Y_1)(t) - T^n(Y_2)(t)\| \leqslant \frac{k^n |t - t_0|^n}{n!} \|Y_1 - Y_2\|_\infty \leqslant \frac{(k\alpha)^n}{n!} \|Y_1 - Y_2\|_\infty$$

As $\frac{(k\alpha)^n}{n!} \to 0$, there is $n \in \mathbb{N}$ such that $\ell = \frac{(k\alpha)^n}{n!} < 1$. Then $T^n$ is an $\ell$-contraction mapping. ◊

We finish with a fact from topology.

**<u>Fact</u>** (a corollary to Picard's fixed point theorem)**.** Let $(C, d)$ be a complete metric space, $\ell < 1$, and $S : C \to C$ a continuous map such that there is $n \in \mathbb{N}$ for which $S^n$ is an $\ell$-contraction mapping.

Then $S$ has a unique fixed point.

Another fact from functional analysis is that $C = C^0\left(\overline{J}, \overline{B}(X_0, r)\right)$ is a complete metric space indeed (for the uniform distance; one may not say norm here, as it is not a vector space). So $T$ has a unique fixed point in $C$, meaning that the Cauchy problem has a unique solution defined on $J$. □

**Remark.** This second proof of the Cauchy-Lipschitz theorem gives another practical way to construct a solution, because for any $Y \in C$ the sequence of iterates $(T^n(Y))$ will converge to the unique solution. This method, known as Picard iterates, is different from Euler's algorithm: it belongs more to functional analysis than to numerical analysis.

**Example.** Return to our toy example: the Cauchy problem $x'(t) = x(t)$ with initial condition $x(0) = 1$. Here, $G(t, x) = x$ is 1-Lipschitz with respect to $x$ (even globally so); $t_0 = 0$ and $x_0 = 1$, so that the functional analysis operator is:

$$T(y) = 1 + \int_0^t y(s) ds$$

Let us compute a few iterates: start with the constant function $y_0 = x_0 =$

1; then:

$$\begin{aligned}
y_1(t) &= T(y_0)(t) = 1 + \int_0^t 1ds = 1 + t;\\
y_2(t) &= T(y_1)(t) = 1 + \int_0^t (1+s)ds = 1 + t + \frac{t^2}{2};\\
&\vdots\\
y_n(t) &= \sum_{k=0}^{n} \frac{t^k}{k!}
\end{aligned}$$

The sequence is the Taylor series of the exponential map.

**Remark.** The same argument gives another proof of the global Cauchy-Lipschitz theorem (just do $r = +\infty$ in the cylinder), which we state again for the reader's convenience.

**Theorem** (Cauchy-Lipschitz: global version)**.** Consider a Cauchy problem with continuous $G$.

Suppose that there is a function $k : I \to \mathbb{R}_{\geqslant 0}$ such that for all $t \in I$, the map $G(t, \cdot)$ is globally $k(t)$-Lipschitz, i.e.:

$$\forall t \in I, \forall (X_1, X_2) \in U^2, \quad \|G(t, X_1) - G(t, X_2)\| \leqslant k(t) \cdot \|X_1 - X_2\|.$$

Then the Cauchy problem has a unique maximal solution *which is global.*

# II.5. The Cauchy-Kovalevskaya theorem

Solutions exist but do not always have an explicit formula. However, one can expand solutions into series.

**Definition.** A function $x : I \to \mathbb{R}^d$ is real analytic around $t_0$ if there are $\eta > 0$ and coefficients $(a_n) \in \mathbb{R}^{\mathbb{N}}$ such that, for all $t \in (t_0 - \eta, t_0 + \eta)$:

$$x(t) = \sum_{n=0}^{\infty} a_n (t - t_0)^n$$

The maximal $\eta$ such that this holds is called the convergence radius.

**Remarks.**

- Sometimes one says that the function is real analytic *at* $t_0$, but this is bad terminology: the condition is local (on a small neighborhood), not punctual (on a singleton).

- A real analytic function is a map which locally equals its Taylor expansion.
- Be extremely careful that although a real analytic function is always $C^\infty$, the converse fails. The classical example is the map:

  $$x \mapsto e^{-1/x^2},$$

  with continuous extension $x(0) = 0$. One then sees that $x$ is infinitely differentiable everywhere including at 0, where $x^{(n)}(0) = 0$. Hence the Taylor series of $f$ at 0 is the null series, which does not give $f$.

  (In particular, $f(0)$ is equal to its Taylor series at 0, but this is not true for any other value: this is why analytic "at" $t_0$ is bad terminology.)
- Interestingly enough, for complex functions the difference between complex analytic and $C^\infty$ vanishes: another name is holomorphic functions; see your class in complex analysis.

There is a multivariable notion of real analytic, but we shall keep things simple.

**Theorem** (Cauchy-Kovalevskaya)**.** Consider a scalar, "autonomous" Cauchy problem $x'(t) = f(x(t))$ for $t \in I$ with initial condition $x(t_0) = x_0$.

Suppose that $f$ is real analytic around $x_0$. Then there exists a unique solution; it is real analytic around $t_0$.

There are very nice proofs relying on complex analysis; we shall follow Cauchy's original argument.

**Proof.** Since $f$ is real analytic around $x_0$, it is $C^\infty$ locally, so it is locally Lipschitz. By the Cauchy-Lipschitz theorem, there exists a unique solution, which we denote by $x(t)$. The theorem now reduces to proving that $x$ itself is analytic around $t_0$.

By standard reductions we may assume $t_0 = 0$ and $x_0 = 0$.

Since $f$ is real analytic around $x_0 = 0$ it has (locally) an expression:

$$f(x) = \sum_n a_n x^n$$

We argue algebraically and entirely forget about convergence issues.

**Definition.** A formal solution is a formal series $y(t) = \sum_n b_n t^n$ such that, formally:

$$\sum_n a_n \left(\sum_k b_k t^k\right)^n = \left(\sum_n b_n t^n\right)'$$

where composition and derivation are in the sense of formal series.

**Remarks.**

- This definition may seem complicated but it is actually straightforward: just follow the standard procedures without caring about convergence issues.
- Here are the relationships between "formal" and "genuine" solutions:
  - if a formal solution has non-zero convergence radius, then it defines a genuine solution;
  - if a genuine solution is analytic, then its series is a formal solution.

**Step 1.** There exists a family of polynomials $(P_n)$, with $P_n$ in $n$ variables and coefficients in $\mathbb{Q}_{\geqslant 0}$, such that:

> if $f(x) = \sum_n a_n x^n$, then there exists a unique *formal* solution to the Cauchy problem $x'(t) = f(t, x(t))$ with initial condition $x(0) = 0$: the formal series $x(t) = \sum_n b_n t^n$ where $b_n = P_n(a_1, \dots, a_n)$,

**Proof**. We are still at the level of algebraic manipulations of series, without discussing convergence. Return to the formal equation, supposing that $\sum_n b_n t^n$ is a formal solution, and write:

$$\sum_n a_n \left(\sum_k b_k t^k\right)^n = \left(\sum_n b_n t^n\right)' = \sum_n n b_n t^{n-1}$$

In the right-hand side, the coefficient of $t^n$ is $(n+1)b_{n+1}$. In the left-hand side, it is a complicated expression involving $a_0, \dots, a_n$ and $b_0, \dots, b_n$ (this is true only since $a_0 = 0$) — but all coefficients are positive integers.

So by induction, $b_n$ is a rational polynomial (with positive coefficients) in $a_1, \dots, a_n$. Let $P_n$ be this polynomial, which does not depend on $f$.

Since we computed formally, the converse also holds: letting $b_n = P_n(a_0, \dots, a_n)$ and $x(t) = \sum b_n t^n$ defines a formal solution, and we have no other choice. ◊

So we know that the Cauchy problem has:

- a unique local solution (this is Cauchy-Lipschitz);
- a unique formal solution (by the above).

Proving that the genuine solution is real analytic around $t_0$ amounts to proving that the formal solution has a non-zero convergence radius.

**Definition.** Let $\sum b_n t^n$ be a formal series, with $b_n \in \mathbb{R}$. A majorant series is a formal series $\sum \beta_n t^n$ such that: $\forall n \in \mathbb{N}, |b_n| \leqslant \beta_n$.

This definition is surprisingly simple; Cauchy's idea is to solve analytically the differential equation associated to a majorant series.

**Step 2.** There are $M, r > 0$ such that $\alpha_n = \frac{M}{r^n}$ defines a majorant series for $\sum a_n x^n$.

**Proof**. By assumption, $f$ is analytic around 0: so there is $x_1 \neq 0$ such that $\sum_n a_n x_1^n$ converges. Then $a_n x_1^n \to 0$; in particular, it is bounded by say $M$. Let $r = |x_1|$ and $\alpha_n = \frac{M}{r^n}$; observe how $|a_n| \leqslant \alpha_n$, $\sum_n \alpha_n x^n$ is a majorant series of $\sum a_n x^n$. ◊

**Step 3.** The map $\varphi(x) = \sum \alpha_n x^n$ is real analytic. The map $\chi(t) = r - \sqrt{r^2 - 2Mrt}$ is the unique solution of $\chi'(t) = \varphi(\chi(t))$ with $\chi(0) = 0$.

**Proof**. Consider the formal series:

$$\varphi(x) = \sum \alpha_n x^n = \sum \frac{M}{r^n} x^n = M \cdot \frac{1}{1 - \frac{x}{r}} = \frac{Mr}{r - x}$$

As we know from analysis, this formal series actually has convergence radius $r$; $\varphi$ is a real analytic function on $(-r, r)$.

Now consider the analytic differential equation:

$$\chi'(t) = \varphi(\chi(t))$$

with initial condition $\chi(0) = 0$. Since $\varphi$ is real analytic around 0, it is $C^\infty$ there, hence also locally Lipschitz. So the Cauchy-Lipschitz theorem applies: there is a unique solution.

But one can check that $\chi(t) = r - \sqrt{r^2 - 2Mrt}$ is locally well-defined, regular, and satisfies both the equation and initial condition. By uniqueness, $\chi$ is the unique solution. ◊

Now $\chi$ is real analytic as we know from analysis; we may write:

$$\chi(t) = \sum \beta_n t^n$$

**Step 4.** $\sum \beta_n t^n$ is a majorant series for $\sum b_n t^n$.

**Proof**.

- On the one hand, $\sum b_n t^n$ is the formal solution to $x'(t) = f(x(t))$ with $x(0) = 0$: hence $b_n = P_n(a_0, \dots, a_n)$.
- On the other hand, $\chi(t)$ is the genuine solution to $\chi'(t) = \varphi(\chi(t))$ with $\chi(0) = 0$. But $\chi$ happens to be analytic, so $\chi$ also is the formal solution to $\chi'(t) = \varphi(\chi(t))$ with $\chi(0) = 0$. In particular $\beta_n = P_n(\alpha_0, \dots, \alpha_n)$.

Now remember that $P_n$ has positive coefficients. Since $|a_n| \leqslant \alpha_n$, we find $|b_n| \leqslant \beta_n$, as desired. ◊

We finish the proof. The formal series $\sum b_n t^n$ is the formal solution to $x'(t) = f(x(t))$ with $x(0) = 0$. But it has a majorant series which converges with non-zero radius. So $\sum b_n t^n$ itself has non-zero radius: as a consequence, the real analytic function defined by $\sum b_n t^n$ is a genuine solution. By uniqueness, it is the function $x(t)$. $\square$

We admit that the theorem still holds for vector equations $X' = G(X(t))$ with $G$ real analytic.

**Examples.**

- Consider the pendulum equation $\theta'' = \sin\theta$. Since sin is analytic around 0, so is the solution.

- The three-body problem in celestial mechanics is given by analytic equations: so the solution is analytic. But in practice, the series giving the solution converges too slowly to be useful.