Università degli Studi di Trento

Department of Industrial Engineering

Mechatronic Systems Simulation

Prof.: Bertolazzi Enrico, Biral Francesco

Course Notes

Matteo Dalle Vedove
matteo.dallevedove@studenti.unitn.it

Academic Year 2021-2022
January 20, 2023

Contents

Part I

Computational Methods

Chapter 1

Laplace Transform

The **Laplace Transform** $\mathscr{L}$ is a powerful operator that allow to express a function $f(t)$ in the domain of the time t to a function $\hat{f}(s)$ expressed in the domain of the **complex variable** s; in a mathematical way the passage from one domain (in this case time) to another (complex variable) is expressed as

$$f(t) \mapsto \hat{f}(s) = \mathscr{L}\{f(t)\}$$

Not all function f can be transformed, and this is due to the existence (or not) of the following integral that is used to calculate the transform of the function:

$$\hat{f}(s) = \int_{0^-}^{\infty} f(t)e^{-st}dt = \lim_{\varepsilon \to 0^+} \lim_{M \to \infty} \int_{-\varepsilon}^{M} f(t)e^{-st}\,dt \tag{1.1}$$

This mathematical tool is very powerful because it can transform **differential equation** (in the domain of the time) **into algebraic equation** (in the domain of s) which are much easier to solve.

> **Note:** This concept can be seen with a logarithm analogy: the product of two number, by the logarithm rule, is easier to calculate because
>
> $$a \cdot b \mapsto \log a + \log b$$
>
> so by this with the **logarithm** you can convert **product** into **sums** that are easier to manipulate.

In mechanical system is often required to solve linear differential equation in order to describe the time response of the system itself: this can be done with analytical techniques (such the *constant variations* method), but can be very tricky to solve, or by using the Laplace transform as follows:

- with the **Laplace transform** the differential equation is converted into an algebraic one;
- by analyzing this equation you can determine the **frequency response** of the system object of study;
- with the **Laplace inverse transform** it's possible to re-convert the solution from the domain of the complex variable s into the domain of time t.

1.1 Transform properties

The Laplace transform has some important properties that can simplify the hand-made calculus operation; the first important thing to keep is mind is that the Laplace transform is a **linear operator**, so for all functions $f(t) \mapsto \hat{f}(s)$ and $g(t) \mapsto \hat{g}(s)$ and real constants $a, b \in \mathbb{R}$ it's true that

$$h(t) := a\,f(t) + b\,g(t) \mapsto \hat{h}(s) := a\,\hat{f}(s) + b\,\hat{g}(s) \tag{1.2}$$

Demonstration 1.1: The linearity property can be demonstrated by applying the Laplace transform equation to the linear combination of two function:

$$\begin{aligned}\mathscr{L}\{a\,f(t)+b\,g(t)\} &= \int_{0^-}^{\infty}\Big(a\,f(t)+b\,g(t)\Big)e^{-st}\,dt\\ &= a\int_{0^-}^{\infty} f(t)e^{-st}\,dt + b\int_{0^-}^{\infty} g(t)e^{-st}\,dt\\ &= a\,\hat{f}(s)+b\,\hat{g}(s)\end{aligned}$$

Note that in order to prove this property we had to use the property of the integral that allowed us to split him in two separate integrals: this can be true in a general case but some function might not satisfy this option, like it can be seen in the example 1.1.

Example 1.1: integrable and non-integrable function

Consider the piecewise defined function (dependent from the parameter n) for $t \geq 0$

$$f_n(t) = \begin{cases} nt & 0 \leq t \leq 1/n \\ 2-nt & 1/n \leq t \leq 2/n \\ 0 & \text{otherwise}\end{cases}$$

By pushing the value n to infinity we can see that the function $f(t) = \lim_{n\to\infty} f_n(t)$ has always a value of zero $\forall t$; by computing the integral (from zero to infinity) it's possible to calculate the area generated by this function that's equal to

$$F_n(t) = \int_0^{\infty} f_n(t)\,dt = \frac{1}{n} \qquad \xrightarrow{n\to\infty} \qquad 0$$

So the area associated to this function, if pushing $n \to \infty$, is zero, as expected because the function is always zero in it's domain, so this might give the (bad) idea that's possible to take out the limit of the integral outside, in the sense that

$$F(t) = \int_0^{\infty} \lim_{n\to\infty} f_n(t)\,dt = \lim_{n\to\infty}\int_0^{\infty} f_n(t)\,dt$$

Considering now another piecewise function g_n defined for $t \geq 0$

$$g_n(t) = \begin{cases} n^2t & 0 \leq t \leq 1/n \\ 2n-n^2t & 1/n \leq t \leq 2/n \\ 0 & \text{otherwise}\end{cases}$$

As in the previous case by computing the limit we see that the function $g = \lim_{n\to\infty} g_n$ is always zero in it's domain; now if we consider the position of the limit we can see that the result of the computed area differs, in fact

$$\lim_{n\to\infty}\int_0^{\infty} g_n(t)\,dt = \frac{\frac{2}{n}n}{2} = 1 \qquad \neq \qquad \int_0^{\infty}\lim_{n\to\infty} g_n(t)\,dt = \int_0^{\infty} 0\,dt = 0$$

So in a general we have to keep in mind that

$$\lim\int \quad \neq \quad \int\lim$$

This concept, for example, must be kept in mind when applying the linearity rule (or in general any other properties).

Another important fact is associated to the **scale change** of the time axes; in particular by stretching/expanding the time axes by a value $a > 0$ of a function $f(t) \mapsto \hat{f}(s)$ it's true that:

$$f(at) \mapsto \frac{1}{a}\hat{f}\left(\frac{s}{a}\right) \tag{1.3}$$

Demonstration 1.2: As in demonstration 1.1, the property of the scale change can be verified by using the definition of the Laplace transform using the change of variables $at = z$ (that means $t = z/a$):

$$\begin{aligned}\mathscr{L}\{f(at)\} &= \int_{0^-}^{\infty} f(at)\,dt \\ &= \int_{0^-}^{\infty} f(z)e^{-sz/a}\,\frac{dz}{a} \\ &= \frac{1}{a}\hat{f}\left(\frac{s}{a}\right)\end{aligned}$$

Other two important properties are related to the **translation** in respect to the s axes as in respect to the t axes (by a coefficient $a > 0$) by using the relation that follows:

$$e^{at}f(t) \mapsto \hat{f}(s-a) \qquad\qquad f(t-a) \mapsto e^{-at}\hat{f}(s) \tag{1.4}$$

Demonstration 1.3: The property of the translation in respect of the s complex variable is done as follows:

$$\begin{aligned}\mathscr{L}\{e^{at}f(t)\} &= \int_{0^-}^{\infty} e^{at}f(t)e^{-st}\,dt = \int_{0^-}^{\infty} f(t)e^{(a-s)t}\,dt \\ &= \hat{f}(s-a)\end{aligned}$$

The demonstration of the translation in respect to time t is a little bit longer and it involves the change of coordinates $z = t - a$:

$$\begin{aligned}\mathscr{L}\{f(t-a)\} &= \int_{0^-}^{\infty} f(t-a)e^{-st}\,dt = \int_{-a}^{\infty} f(z)e^{-s(z-a)}\,dz \\ &= e^{-sa}\cancel{\int_{-a}^{0^-} f(z)e^{-sz}\,dz} + e^{-sa}\int_{0^-}^{\infty} f(z)e^{-sz}\,dz \\ &= e^{-as}\hat{f}(s)\end{aligned}$$

During the decomposition it's possible to cancel out the first integral (second step) because in general we consider function $f(t)$ that are always zero for $t < 0$, so by computing the integral that always going to be zero.

1.2 Existence of the transform

Note that not all function can be transformed using the Laplace operator $\mathscr{L}$; if, for example, we choose the function $f(t) = e^{t^2}$, by applying the Laplace transform to this expression we get the following integral

$$\mathscr{L}\left\{e^{t^2}\right\} = \int_{0^-}^{\infty} e^{t^2-st}\,dt = \int_{0^-}^{T} e^{(t-s)t}\,dt + \int_{T}^{\infty} e^{(t-s)t}\,dt$$

If we consider a costant $T > \mathrm{Re}\,(s)$ we can notice that the second integral $\int_T^{\infty} e^{(t-s)t}\,dt$ is not convergent for any variable $s \in \mathbb{C}$.

Demonstration 1.4: The full mathematical demonstration of this exercise can be seen as follows. Given the function $f(t) = e^{t^2}$, if it's transform exists (and we call it $\hat{f}(s)$) should be equal to

$$\hat{f}(s) = \lim_{M\to\infty} \int_{0^-}^{M} e^{t(t-s)}\, dt$$

This is a complex integral: if we consider that s is a complex variable, that means that can be expressed in the form $s = a + ib$ (where $i = \sqrt{-1}$ and $a, b \in \mathbb{R}$). Substituting this relation in the formal expression of the transform with get the argument of the integral that is equal to $e^{t(t-a-ib)} = e^{t(t-a)}e^{-ib}$. Using the De-Moire formula[a] it's possible to re-write the previous integral as

$$\lim_{M\to\infty} \int_{0^-}^{M} e^{t(t-a)} \Big(\cos(tb) - i\sin(tb) \Big)\, dt$$

$$\Rightarrow \qquad \text{Re}: \lim_{M\to\infty} \int_{0^-}^{M} e^{t(t-a)} \cos(tb)\, dt \qquad \text{Im}: \lim_{M\to\infty} \int_{0^-}^{M} e^{t(t-a)} \sin(tb)\, dt$$

Considering the real part of the integral we can see that it's made of a cosine function that multiplies an exponential (that's monotonically ascendant function). Considering now a point M on the time axes where the cosine start increasing from zero and a consequent point with relative distance ΔM we get that the integral up to M is less then the area up to $M + \Delta M$ and so with that we can see some information:

$$\begin{aligned} \int_0^{M+\Delta M} e^{t(t-a)} \cos(tb)\, dt &\geq \int_0^{M} e^{t(t-a)} \cos(tb)\, dt + \int_M^{M+\Delta M} e^{M(M-a)} \cos(tb)\, dt \\ &\geq \int_0^{M} e^{t(t-a)} \cos(tb)\, dt + e^{M(M-a)} \int_M^{M+\Delta M} \cos(tb)\, dt \end{aligned}$$

this inequality is true because the exponential keeps growing and so by considering it constant in the range $[M, M + \Delta M]$ will give an integral with less value. This integral has a residual part that's not going to zero, and thus the limit doesn't converge. We have in fact to keep in mind that an improper integral exists if and only if the next equation is verified:

$$\lim_{a,b\to\infty} \int_a^b f(t)\, dt = 0 \tag{1.5}$$

[a] $e^{A+iB} = e^A e^{iB} = e^A(\cos B + i \sin B)$.

Exponential order functions A function $f(t)$ can be seen as of **exponential order** function if only happens that $|f(t)| \leq Me^{Nt}$ after a certain time $t \geq T$ (and $M, N \in \mathbb{R}$ are two constants); in this case (if $f(t)$ is a exponential order function) it's true that **Laplace transform always exists** (in at least a part of the complex plane).

Demonstration 1.5: To demonstrate that a exponential order function has always a Laplace transform by considering only the residual part of the integral, so ranging not from 0^-, but from T to infinity. In particular we can verify the all the following inequalities are correct:

$$\begin{aligned} \left| \int_0^T f(t)e^{-st}\, dt \right| \leq \left| f(t)e^{-st} \right|\, dt &\leq \int_0^{\infty} |f(t)| \left| e^{-st} \right|\, dt \\ &\leq \int_0^{\infty} Me^{Nt} \left| e^{-st} \right|\, dt \end{aligned}$$

By using the De-Moire formula it's possible to rewrite the $|e^{-st}|$ term that end up resulting

$$\begin{aligned} \left| e^{-st} \right| &= \left| e^{-at} \big(\cos(bt) + i\sin(bt) \big) \right| = e^{-at} \left| \cos(bt) + i\sin(bt) \right| \\ &= e^{-at} \cancel{\sqrt{\cos^2(bt) + \sin^2(bt)}} \end{aligned}$$

At this point we can continue in the analysis of the inequalities keeping in mind that now on $a = \mathrm{Re}(s)$:

$$\left| \int_0^T f(t)e^{-st}\,dt \right| \leq \int_T^\infty Me^{Nt}e^{-at}\,dt = M\int_T^\infty e^{(N-a)t}\,dt$$

At this point if $a > N$ we see that the residual $e^{(N-a)t} \xrightarrow{t\to\infty} 0$ goes to zero, hence the integral should have a limited value.

By following the definition it's also possible to observe that the **domain** of the variable s for the Laplace transform of an exponential order function is **at least** $\forall s = a + ib \in \mathbb{C}$ such that $a = \mathrm{Re}(s) \geq N$.

1.3 Usage of the Laplace transform

Imagine the problem of solving the following differential equation problem

$$\begin{cases} y'(t) = ay(t) + t \\ y(0) = 1 \end{cases}$$

Solving this equation in the domain of time can be difficult in general, and that's why we introduced the Laplace transform and the analysis in the domain of the complex variable s. By computing the Laplace transform on the differential equation we get the expression

$$\begin{aligned} \mathscr{L}\left\{y'(t)\right\}(s) &= \mathscr{L}\left\{ay(t) + t\right\}(s) \\ &= a\mathscr{L}\left\{y(t)\right\}(s) + \mathscr{L}\left\{t\right\}(s) \end{aligned}$$

At this point we have to describe an important **Laplace transform property** that states that the transform of the first derivative $f'(t)$ of a function $f(t) \mapsto \hat{f}(s)$ is equal to the transform $\hat{f}(s)$ multiplied by the complex variable s:

$$f'(t) \mapsto s\,\hat{f}(s) - f(0^+) \tag{1.6}$$

where $f(0^+)$ is the value of the function $f(t)$ at the origin of the time axes and, de-facto, represents the initial (Cauchy) condition of the ordinary differential equation problem.

This property allow us to solve **differential equation** in the time domain as **algebraic equations** in the complex variable world; returning back to the initial problem we can rewrite the differential equation as

$$s\,\hat{y}(s) - \underbrace{1}_{y(0^+)} = a\,\hat{y}(s) + \underbrace{\frac{1}{s^2}}_{\mathscr{L}\{t\}} \qquad \Rightarrow \quad \hat{y}(s) = \frac{1 + \frac{1}{s^2}}{s - a}$$

Note: To do the calculation it's useful to refer to the property table (A.1) and the common functions transforms (table A.1) on page 131.

Example 1.2: system of ordinary differential equation

Consider the following system of 2 ordinary differential equation subjected to their initial condition as follows:

$$\begin{cases} x'(t) = y(t) + 1 \\ y'(t) = x(t) + \cos t \\ x(0) = 1 \\ y(0) = 1 \end{cases}$$

To solve this particular problem we can apply the rules of the Laplace transform to determine the

system of 2 equation in the domain of the complex variable s as follows:

$$i) \qquad \underbrace{s\tilde{x}(s) - x(0)}_{\mathscr{L}\{x'(t)\}} = \tilde{y}(s) + \frac{1}{s}$$

$$ii) \qquad \underbrace{s\tilde{y}(s) - y(0)}_{\mathscr{L}\{y'(t)\}} = \tilde{x}(s) + \frac{1}{1+s^2}$$

Knowing the initial condition of the system it's possible to explicit the system of the function $\tilde{x}, \tilde{y}$ by using a matrix notation arriving to the following result:

$$\underbrace{\begin{bmatrix} s & -1 \\ -1 & s \end{bmatrix}}_{A(s)} \begin{pmatrix} \tilde{x}(s) \\ \tilde{y}(s) \end{pmatrix} = \begin{pmatrix} 1 + \frac{1}{s} \\ \frac{1}{1+s^2} \end{pmatrix}$$

The explicit value of the variables $\tilde{x}, \tilde{y}$ can be calculated by inverting the matrix $A(s)$ arriving to the following results:

$$\tilde{x}(s) = \frac{s + 1 - \frac{1}{1+s}}{s^2 - 1} \qquad \tilde{y}(s) = \frac{\frac{s}{1+s} - 1 - \frac{1}{s}}{s^2 - 1}$$

Example 1.3: computation of the Laplace transform

Given the ordinary differential equation

$$\begin{cases} x'(t) + x(t) = \cos t \\ x(0) = 2 \end{cases}$$

the solution can be found in the Laplace domain by using the operator $\mathscr{L}$ on the ODE considering the known transforms, in particular

$$\mathscr{L}\{x(t)\} = \hat{x}(s) \qquad \mathscr{L}\{x'(t)\} = s\hat{x}(s) - x(0) \qquad \mathscr{L}\{\cos(t)\} = \frac{s}{s^2+1}$$

$$\mathscr{L}\{x'(t) + x(t) = \cos t\} \quad \mapsto \quad s\hat{x}(s) - 2 + \hat{x}(s) = \frac{s}{s^2+1}$$

Solving this expression for $\hat{x}(s)$ gives

$$\hat{x}(s) = \frac{s}{(s+1)(s^2+1)} + \frac{2}{s+1} = \frac{2s^2 + s + 2}{(s+1)(s^2+1)}$$

1.4 Inversion of the Laplace transform: partial fraction expansion

The **inversion** of a Laplace transform is the step that allow to *transport* the algebraic solution found in the s domain into a time description, and so it's represented by the operator $\mathscr{L}^{-1}$.

An important thing to notice is that while dealing with (systems of) ordinary differential equation is that the result of the transform (such as $\hat{x}(s)$) is usually expressed as a rational polynomial in the form $P(s)/Q(s)$ (where P, Q are so two polynomial with real coefficients in the variable s). It's important also to note that, very often in real application, the degree ∂P of the numerator is less than the degree of the denominator ∂Q. When this doesn't *naturally* happen, the step to follow is to use the polynomial division: considering in fact that $P(s)$ in this case can always be rewritten as $T(s)Q(s) + R(s)$, it's easy

to see that

$$\frac{P(s)}{Q(s)} = \frac{T(s)Q(s) + R(s)}{Q(s)} = T(s) + \frac{R(s)}{Q(s)}$$

The part represented by the polynomial $T(s)$ that has real coefficients is associated to the transform of Dirac pulses function $\delta(t)$: we can in fact note

$$\begin{aligned} \mathscr{L}\left\{\delta(t)\right\}(s) &= \int_{0^-}^{\infty} \delta(t)e^{-st}\,dt = e^{-st}\Big|_{t=0} \\ &= 1 \\ \mathscr{L}\left\{\delta'(t)\right\}(s) &= -\int_{0^-}^{\infty} \delta'(t)e^{-st}\,dt = -\frac{d}{dt}\left(e^{-st}\right)\Big|_{t=0} = se^{-st}\Big|_{t=0} \\ &= s \end{aligned} \tag{1.7}$$

Assumed that we now have a transform solution that's a simplified rational polynomial $P(s)/Q(s)$ such that $\partial P < \partial Q$, we can try to compute it's inverse, so the same function but expressed in the time domain. In practise this is done by using the **partial fraction expansion** of the rational polynomial: this allows to re-state the solution as a combination of simpler *elements* that can be easily inverted by simply looking a the Laplace table.

In order to properly do a partial fraction expansion is important to re-write the denominator of the polynomial ratio in a factorised form such as

$$\frac{P(s)}{Q(s)} = \frac{b_0 + b_1 s + b_2 s^2 + \cdots + b_m s^m}{(s-p_1)^{m_1}(s-p_1)^{m_2}\ldots(s-p_1)^{m_n}}$$

where p_i are the roots of the denominator everyone having a molteplicity m_i. Depending on the molteplicity of the root and their kind (real or conjugated complex) the procedure to accomplish a partial fraction example can be different. In the following paragraph each possibility will be presented with an example.

1.4.1 Real roots

Single real root Let's consider the following rational polynomial with no multiple roots:

$$G(s) = \frac{s}{(s-1)(s+3)(s-4)}$$

It's factorization is based on determining the coefficients α_i such that

$$i): \qquad \frac{\alpha_1}{s-1} + \frac{\alpha_2}{s+3} + \frac{\alpha_3}{s-4} = \frac{s}{(s-1)(s+3)(s-4)}$$

This can be accomplished in two way: the first one is to compute the some of the 3 *basic* element in a parametric form depending on α_i and then solving the linear system in order to determine the correct parameters to satisfy the equality. A second (smarter) way is instead considering that each coefficient can be determined as

$$\alpha_i = \lim_{s \to p_i} (s - p_i)G(s)$$

Considering the example, to compute the coefficient α_1 associated to the root $p = 1$ we can multiply the relation $i)$ with a factor $(s-1)$: this gives us the equation

$$\frac{s}{\cancel{(s-1)}(s+3)(s-4)}\cancel{(s-1)} = \frac{\alpha_1}{\cancel{s-1}}\cancel{(s-1)} + (s-1)\left(\frac{\alpha_2}{s+3} + \frac{\alpha_3}{s-4}\right)$$

Evaluating this expression at the point $s = 1$ will result in an equation whose only unknown coefficient is α_1 (because α_2, α_3 are *eliminated* to the multiplication with $s - 1$ that goes to zero), and so

$$\xrightarrow{s=1} \quad \alpha_1 = \frac{1}{(1+3)(1-4)} = -\frac{1}{12}$$

Repeating this process also for the other two roots $(-3, 4)$ we retrieve the coefficients $\alpha_2 = -\frac{3}{28}$ and $\alpha_3 = \frac{4}{21}$: with that is possible to verify that

$$-\frac{1}{12}\frac{1}{s-1} - \frac{3}{28}\frac{1}{s+3} + \frac{4}{21}\frac{1}{s-4} = \frac{s}{(s-1)(s+3)(s-4)}$$

Each basic element can be easily transform using the Laplace table, and in particular

$$\mathscr{L}^{-1}\left\{G(s)\right\}(t) \; : \quad -\frac{1}{12}e^{-t} - \frac{3}{28}e^{3t} + \frac{4}{21}e^{-4t}$$

Multiple real roots Let's consider now a case in which the rational polynomial has multiple roots such as

$$G(s) = \frac{1+s}{(s-4)^4}$$

In this case the partial fraction has to be made not with 4 equals element with denominator $s - 4$, but instead consider four increasing exponential terms as follows:

$$i) \qquad \frac{\alpha_1}{s-4} + \frac{\alpha_2}{(s-4)^2} + \frac{\alpha_3}{(s-4)^3} + \frac{\alpha_4}{(s-4)^4} = \frac{1+s}{(s-4)^4}$$

This problem can be solved by summing up all the elements and solving the linear system in respect to α_i, but a smarter way is to consider that the element a_{n-k} (where n is the number of the multiple root, in this case 4) can be expressed as $\frac{1}{k!}\lim_{s\to p}\frac{d^k}{ds^k}\left[(s-p)^n G(s)\right]$; this process is easier to see considering the stated example. The first thing is to multiply $i)$ with the polynomial $(s-4)^4$ arriving to the relation

$$1 + s = \alpha_1(s-4)^3 + \alpha_2(s-4)^2 + \alpha_3(s-4) + \alpha_4$$

Evaluating this expression for $s = 4$ determines the coefficient $\alpha_4 = 5$; in order to determine the other coefficient is to use the proposed *scheme* where each time we differentiate (in respect to s) the previous equation and evaluate at the point $s = 4$:

$$\begin{aligned}
&\downarrow d/ds \\
&1 = 3\alpha_1(s-4)^2 + 2\alpha_2(s-4) + \alpha_3 && \xrightarrow{s=4} \quad \alpha_3 = 1 \\
&\downarrow d/ds \\
&0 = 6\alpha_1(s-4) + 2\alpha_2 && \xrightarrow{s=4} \quad \alpha_2 = 0 \\
&\downarrow d/ds \\
&0 = 6\alpha_1(s-4) && \xrightarrow{s=4} \quad \alpha_1 = 0
\end{aligned}$$

Having determined the coefficients α_i associated to the partial fraction expansion it's possible to compute the inverse transform of the original $G(s)$ that's

$$\mathscr{L}^{-1}\left\{G(s)\right\} = \mathscr{L}^{-1}\left\{\frac{1}{(s-4)^3}\right\} + 5\mathscr{L}^{-1}\left\{\frac{1}{(s-4)^4}\right\} = \frac{1}{2}e^{4t}t^2 + 5\frac{1}{6}e^{4t}t^3$$

Example 1.4: partial fraction expansion with different roots, one which is multiple

Let's consider the following rational polynomial $G(s)$ that's already been split in the factor for the partial expansion:

$$G(s) := \frac{1+s}{(s-1)(s+2)^3} = \frac{A}{s-1} + \frac{B}{s+2} + \frac{C}{(s+2)^2} + \frac{D}{(s+2)^3}$$

In order to calculate the first coefficient A related to the root $s = 1$ we simply need to multiply the last equation with a factor $(s-1)(s+2)^2$ and then evaluate the result in respect to $s = 1$:

$$\begin{aligned} 1+s &= A(s+2)^3 + (s+1)\Big[B(s+2)^2 + C(s+2) + D\Big] \\ \xrightarrow{s=1} \qquad 2 &= A3^3 \qquad \Rightarrow \quad A = \frac{2}{27} \end{aligned}$$

The next step is to calculate the remaining coefficients (B, C, D) associated to the multiple root $s = -2$, however with this equation as now stated we cannot apply the method shown; in order to create a proper equation we need to **deflate** it, so by expanding the $(s+2)^3$ associated to the A coefficient and moving it on the left hand side:

$$\begin{aligned} 1+s-\frac{2}{27}\Big(s^2+6s^2+12s+8\Big) &= (s+1)\Big[B(s+2)^2 + C(s+2) + D\Big] \\ \frac{1}{27}\Big(-2s^3-12s^2+35s+11\Big) &= \end{aligned}$$

At this point we can divide both sides of the equation by a factor $s-1$ (if this cannot be done on the left hand side, this means that an error in calculation has been done), and so by performing the polynomial division we get the equation

$$\frac{1}{27}\Big(-2s^2-14s-11\Big) = B(s+2)^2 + C(s+2) + D$$

From this point onward it's possible to use the method shown to calculate the coefficient of the multiple roots; starting evaluating the previous expression at point $s = -2$ we get the value $D = \frac{-8+28-11}{27} = \frac{1}{3}$. Continuing with differentiation we get

$$\begin{aligned} &\downarrow d/ds \\ \frac{-4s-14}{27} &= 2B(s+2) + C \qquad \xrightarrow{s=-2} \quad C = -\frac{2}{9} \\ &\downarrow d/ds \\ \frac{-4}{27} &= 2B \qquad \xrightarrow{s=-2} \quad B = -\frac{2}{27} \end{aligned}$$

To end the inversion of the starting transform $G(s)$ we can see that

$$\begin{aligned} \mathscr{L}^{-1}\left\{G(s)\right\}(t) &= \frac{2}{27}\mathscr{L}\left\{\frac{1}{s-1}\right\} - \frac{2}{27}\mathscr{L}\left\{\frac{1}{s+2}\right\} - \frac{2}{9}\mathscr{L}\left\{\frac{1}{(s+2)^2}\right\} + \frac{1}{3}\mathscr{L}\left\{\frac{1}{(s+2)^3}\right\} \\ &= \frac{2}{27}e^{t} - \frac{2}{27}e^{-2t} - \frac{2}{9}e^{-2t}t + \frac{1}{3}\cdot\frac{1}{2}e^{-2t}t^2 \end{aligned}$$

1.4.2 Complex roots

Dealing with complex roots on the denominator $Q(s)$ of the solution of the ordinary differential equation has an higher computational difficulty in respect to the real one. An observation that can be done

is that if the complex value $\alpha = a + ib \in \mathbb{C}$ is a root of the denominator, so that $Q(\alpha) = 0$, than also it's conjugate $\alpha^* = a - ib$ it's a root. Considering in fact that the polynomial denominator $Q(s)$, evaluated in the point $s = \alpha$, can be expressed as

$$Q(\alpha) = q_0 + \alpha q_1 + \alpha^2 q_2 + \cdots + \alpha^n q_n = 0$$

where q_i are real coefficients. At this point we have to consider the following property of the conjugate that are true for all couple of complex variables α, β:

$$\begin{aligned} &i) \qquad \left(\alpha + \beta\right)^* = \alpha^* + \beta^* \\ &ii) \qquad \left(\alpha\beta\right)^* = \alpha^*\beta^* \\ &iii) \qquad \alpha\alpha^* = |\alpha|^2 \end{aligned} \tag{1.8}$$

Now known that $Q(\alpha) = 0$, than also the conjugate $Q^*(\alpha)$ should be equal to zero (because $0^* = 0$) and so by applying the properties we can see that

$$\begin{aligned} 0 = Q(\alpha) = Q^*(\alpha) &= \left(q_0 + \alpha q_1 + \alpha^2 q_2 + \cdots + \alpha^n q_n\right)^* \\ &= q_0^* + \alpha^* q_1^* + \alpha^{*2} q_2^* + \cdots + \alpha^{*n} q_n^* \\ &= q_0 + \alpha^* q_1 + \alpha^{*2} q_2 + \cdots + \alpha^{*n} q_n = Q\left(\alpha^*\right) \end{aligned}$$

This verifies that if α is a root of $Q(s)$, than also it's conjugate α^* is a root of the same polynomial and so we can rewrite the denominator in the form

$$Q(s) = \ldots \underbrace{\left(s - \alpha\right)\left(s - \alpha^*\right)}_{\text{conj. roots}} \ldots$$

Single conjugated complex roots Let's now assume that we have a rational polynomial that present two conjugated complex roots expressed in the form

$$\frac{P(s)}{Q(s)} = \frac{P(s)}{(s-\alpha)(s-\alpha^*)} = \frac{A}{s-\alpha} + \frac{B}{s-\alpha^*}$$

By trying to apply the partial fraction expansion as for the real root previously described, we can demonstrate that the two coefficients A and B are conjugated, so such that $B = A^*$. In fact by summing the two elements we can see that

$$\frac{A}{s-\alpha} + \frac{B}{s-\alpha^*} = \frac{A(s-\alpha^*) + B(s-\alpha)}{(s-\alpha)(s-\alpha^*)} = \frac{(A+B)s - (A\alpha^* + B\alpha)}{(s-\alpha)(s-\alpha^*)}$$

By expanding the definition of the complex variable (so expanding $A = a + ib$, $B = c + id$...) and considering that the polynomial $P(s)$ at the numerator has only real coefficients it's possible to see that, in order to verify the equality, it must be $a = c$ and $b = -d$, and so the two complex values A and B must be conjugated.

Proven that $B = A^*$ now the problem is to find such complex value A in order to satisfy the partial fraction expansion; in order to do so we can try to use the same approach shown for real roots by multiplying both members of the equation by $Q(s)$ and then evaluating the result for $s = \alpha$:

$$\frac{P(s)}{(s-\alpha)(s-\alpha^*)} = \frac{A}{s-\alpha} + \frac{B}{s-\alpha^*} \qquad \xrightarrow[Q(s)]{\times} \qquad P(s) = A(s-\alpha^*) + A^*(s-\alpha)$$

$$\Rightarrow \qquad P(\alpha) = A(\alpha - \alpha^*) + \cancel{A^*(\alpha - \alpha)}$$

Considering now that $\alpha - \alpha^*$ conresponds to a purely imaginary number equals 2 times the imaginary part of the root α, we can invert that relation in order to explicitly get A:

$$A = \frac{P(\alpha)}{2i\text{Im}\left(\alpha\right)} = -i\frac{P(\alpha)}{2\text{Im}\left(\alpha\right)}$$

Let's now try to consider a real case scenario where we want to deal with the inversion of the following rational polynomial (that's consequently expanded):

$$\frac{s}{\big(s-(1+i)\big)\big(s-(1-i)\big)} = \frac{A}{s-(1+i)} + \frac{A^*}{s-(1-i)}$$

We can now multiply both members of the equation by the denominator $Q(s)$ and then evaluate the result for the point $s = 1+i$ (root of the denominator):

$$s = A\big(s-(1-i)\big) + A^*\big(s-(1+i)\big) \qquad \xrightarrow{s=1+i} \qquad 1+i = A\big(1+i-1+i\big)$$

$$\Rightarrow \qquad A = \frac{1+i}{2i} = \frac{1}{2}(1-i)$$

With this value retrieved we can now state that

$$\frac{s}{\big(s-(1+i)\big)\big(s-(1-i)\big)} = \frac{1}{2}\left(\frac{1-i}{s-(1+i)} + \frac{1+i}{s-(1-i)}\right)$$

Unfortunately watching the Laplace table we can see there's no possible match between the known transform (where the numerator is expressed as a pure real number) and the complex coefficient retrieved. But an interesting fact that we can note is that by *merging* the two components we can get a rational polynomial with real coefficient: in a general way we can in fact state that

$$\begin{aligned}\frac{A}{s-\alpha} + \frac{A^*}{s-\alpha^*} &= \frac{A(s-\alpha^*) + A^*(s-\alpha)}{(s-\alpha)(s-\alpha^*)} \\ &= \frac{s\big(A+A^*\big) - A\alpha^* - A^*\alpha}{s^2 - s\big(\alpha+\alpha^*\big) + |\alpha|^2}\end{aligned}$$

This rational polynomial can be matched with the following transform:

$$\begin{aligned}\mathscr{L}\left\{Ce^{at}\cos(\omega t) + De^{at}\sin(\Omega t)\right\} &= C\frac{s-a}{(s-a)^2+\omega^2} + D\frac{\omega}{(s-a)^2+\omega^2} \\ &= \frac{Cs - \big(D\omega - Ca\big)}{(s-a)^2+\omega^2}\end{aligned}$$

Given now the values of α, A, the problem is to find the expression of the coefficients a, ω, C and D; in order to to so we can perform two steps:

1. match the denominator in order to find a and ω; in order to compare the denominator it's useful to expand the denominator $(s-a)^2+\omega^2 = s^2 - 2sa + a^2 + \omega^2$. At this point by relating the function we can see that

$$\cancel{s^2} - s\big(\alpha+\alpha^*\big) + |\alpha|^2 = \cancel{s^2} - 2sa + a^2 + \omega^2$$

$$\Rightarrow \quad \begin{cases} -2sa = -s\big(\alpha+\alpha^*\big) \\ \alpha^2 + \omega^2 = |\alpha|^2 \end{cases} \qquad \Rightarrow \quad a = \frac{\alpha+\alpha^*}{2} = \mathrm{Re}\,(\alpha)$$

$$\Rightarrow \quad \omega = \sqrt{|\alpha|^2 - a^2} = \sqrt{|\alpha|^2 - \mathrm{Re}^2\alpha} = \sqrt{\mathrm{Im}^2\alpha} = \pm\mathrm{Im}\,(\alpha)$$

2. at this point, known a and ω, it's possible to match the numerator in order to get the coefficients C and D:

$$s\big(A+A^*\big) - A\alpha^* - A^*\alpha = Cs - \big(D\omega - Ca\big)$$

$$\Rightarrow \quad \begin{cases} A + A^* = C \\ -A\alpha^* - A^*\alpha = D\omega - Ca \end{cases} \qquad \Rightarrow \quad C = A + A^* = 2\mathrm{Re}\,(\alpha)$$

$$\Rightarrow \qquad D = \frac{Ca - A\alpha^* - A^*\alpha}{\omega}$$

Example 1.5: inversion of a transform with two conjugated complex roots

Let's consider the following rational polynomial that has to be inverted:

$$\frac{s+1}{s^2+s+10}$$

At this point we can verify that the roots of the denominator are complex conjugated by simply applying the quadratic formula to retrieve them $s_{1,2} = \frac{-1\pm\sqrt{-9}}{2}$. At this point the problem of the inversion if finding the parameters a, ω, C, D such that

$$\frac{s+1}{s^2+s+10} = C\frac{s-a}{(s-a)^2+\omega^2} + D\frac{\omega}{(s-a)^2+\omega^2}$$

By comparing the denominators we can observe that $-2sa = s$ and so $a = -\frac{1}{2}$ and so

$$a^2+\omega^2 = 10 \qquad \Rightarrow \quad \omega = \sqrt{10-a^2} = \frac{\sqrt{39}}{2}$$

At this point it's possible to compare the denominator (considering that now a, ω are known) that's lead to

$$s+1 = C\,(s-a) + D\omega \qquad \Rightarrow \quad C = 1 \quad , \quad D = \frac{1+Ca}{\omega} = \frac{1}{\sqrt{39}}$$

Retrieved all the coefficient we can explicitly state the inversion of the original rational polynomial:

$$\mathscr{L}\left\{\frac{s+1}{s^2+s+10}\right\}(t) = e^{-\frac{1}{2}t}\left[\cos\left(\frac{\sqrt{39}}{2}t\right) + \frac{1}{\sqrt{39}}\sin\left(\frac{\sqrt{39}}{2}t\right)\right]$$

Multiple complex roots Dealing with multiple complex roots requires even more computational steps; let's consider the generic case of a multiple complex root of multiplicity of 2 that's already expanded:

$$\frac{P(s)}{(s-\alpha)^2(s-\alpha^*)^2} = \frac{A}{s-\alpha} + \frac{B}{(s-\alpha)^2} + \frac{A^*}{s-\alpha^*} + \frac{B^*}{(s-\alpha^*)^2}$$

As in all the previously cases in order to compute the first parameter B it's possible to multiply both sides of the equation by the denominator $Q(s) = (s-\alpha)^2(s-\alpha^*)^2$ and then evaluate the relation at the root $s = \alpha$:

$$P(s) = A(s-\alpha)(s-\alpha^*)^2 + B(s-\alpha^*)^2 + A^*(s-\alpha^*)(s-\alpha)^2 + B^*(s-\alpha)^2$$

$$\Rightarrow \quad P(\alpha) = B\underbrace{(\alpha-\alpha^*)}_{=2\mathrm{Im}(\alpha)}{}^2 \qquad \Rightarrow \quad B = \frac{P(\alpha)}{4\mathrm{Im}^2(\alpha)}$$

Now in order to calculate the second parameter A of the relation we have to deflate te elements associated to B and then, as for the multiple real roots, compute the derivative in respect to the variable s:

$$\xrightarrow{\text{deflation}} \quad P(s) - B(s-\alpha^*)^2 - B^*(s-\alpha)^2 = A(s-\alpha)(s-\alpha^*)^2 + A^*(s-\alpha^*)(s-\alpha)^2$$

$$\downarrow \; d/ds$$

$$P'(s) - 2B(s-\alpha^*) - 2B^*(s-\alpha) = A(s-\alpha^*)\left[s-\alpha^* + s(s-\alpha)\right] + A^*(s-\alpha)\left[s-\alpha+2(s-\alpha^*)\right]$$

By now evaluating the expression at the point $s = \alpha$ it's possible to have a clear definition of the complex variable A:

$$P'(\alpha) - 2B\big(s - \alpha^*\big) = A\big(\alpha - \alpha^*\big)^2 \qquad \Rightarrow \quad A = \frac{P'(\alpha) - 2B\big(s - \alpha^*\big)}{4\mathrm{Im}^2(\alpha)}$$

Example 1.6: inversion of a transform with multiple complex roots

Let's consider the following rational transform that has to be inverted:

$$\frac{s^2}{\big(s^2 + s + 1\big)^2}$$

The first thing to notice that the multiple roots are complex because calculating their value by using the quadratic formula it happens that

$$s_{1,2} = \frac{-1 \pm \sqrt{1-4}}{2} = -\frac{1}{2} \pm i\frac{\sqrt{3}}{2}$$

For sake of semplicity of representation of the equation, from now on these roots value will be reported as $z = -\frac{1}{2} + \frac{\sqrt{3}}{2}$ and $z^* = -\frac{1}{2} - \frac{\sqrt{3}}{2}$. At this point it's possible to express the parametric partial fraction expansion of the transform as

$$\frac{s^2}{\big(s^2 + s + 1\big)^2} = \frac{A}{s - z} + \frac{A^*}{s - z^*} + \frac{B}{(s - z)^2} + \frac{B^*}{(s - z^*)^2}$$

In order to retrieve the value of the coefficient B we multiply both the sides by the factor $Q(s) = (s^2 + s + 1)^2 = (s - z)^2(s - z^*)^2$ that, evaluated at the point $s = z$, will give the expression

$$z^2 = B\underbrace{\big(z - z^*\big)}_{i\sqrt{3}}{}^2 \qquad \Rightarrow \quad B = -\frac{z^2}{3} = -\frac{\left(-\frac{1}{2} + i\frac{\sqrt{3}}{2}\right)^2}{3} = -\frac{1}{6} - i\frac{\sqrt{3}}{6}$$

By deflating the term presenting the parameter B just calculate we can get the relation

$$s^2 - B(s - z^*)^2 - B^*(s - z)^2 = A(s - z)(s - z^*)^2 + A^*(s - z^*)(s - z)^2$$

$$\xrightarrow{d/ds} \quad 2s - 2B(s - z^*) - 2B^*(s - z) = A(s - z^*)^2 + 2(A + A^*)(s - z)(s - z^*) + A^*(s - z)^2$$

Having computed the derivative of the relation in respect to the variable s it's possible to compute the parameter A just by evaluating the result in the point $s = z$:

$$2z - 2B\big(z - z^*\big) = A\big(z - z^*\big)^2 \qquad \Rightarrow \quad A = -\frac{2z - 4iB\mathrm{Im}\,(z)}{4\mathrm{Im}^2(z)} = -i\frac{2}{3\sqrt{3}}$$

Parametric solution for multiple complex roots To solve the partial fraction expansion of multiple complex roots it's possible to use a parametric formula where the numerator are defined as polynomials A, B in a generic variable t; in particular for the double complex root the expansion can be expressed as

$$\frac{P(s)}{\big(s^2 - as + t\big)^2 Q(s)} = \frac{A'(t)s + B'(t)}{s^2 - as + t} - \frac{A(t)s + B(t)}{\big(s^2 - as + t\big)^2} + \frac{q_t(s,t)}{Q(s)} \tag{1.9}$$

where the term $q_t(s,t)$ represent the derivative $\partial q/\partial t$ of the remaining term on the numerator after the partial fraction expansion. If the polynomial has 3 complex roots than the parametric decomposition

becomes

$$2\frac{P(s)}{\left(s^2-2as+t\right)^3Q(s)} = 2\frac{A(t)s+B(t)}{\left(s^2-2as+t\right)^3} - 2\frac{A'(t)s+B'(t)}{\left(s^2-2as+t\right)^2} + \frac{A''(t)s+B''(t)}{s^2-as+t} + \frac{q_{tt}(s,t)}{Q(s)} \tag{1.10}$$

where $q_{tt} = \partial^2 q/\partial t^2$.

This process can be better seen by considering a numerical example like the one that follows:

$$\frac{s^2+1}{(s-1)(s^2-2s+2)^3} = \frac{A}{s-1} + \frac{B_1 s+B_2}{(s^2-2s+2)^3} + \frac{C_1 s+C_2}{(s^2-2s+2)^2} + \frac{D_1 s+D_2}{s^2-2s+2}$$

The first coefficient A of the partial fraction expansion (associated to a single real root) can be determined easily with the methods previously described and so

$$A = \left.\frac{s^2+1}{(s^2-2s+2)^3}\right|_{s=1} = 2$$

At this point it's important to acknowledge hot compute the parametric functions A, B, q; in particular this can be done by writing the following relation:

$$\frac{s^2+1}{(s-1)(s^2-2s+2)} = \frac{A(t)s+B(t)}{s^2-2s+t} + \frac{q(t)}{s-1}$$

We can note that the left hand side is equal to the original rational polynomial with the only difference that the term s^2-2s+2 is now considered once; the right hand side is instead the *correct* partial fraction expansion of the other side with the difference that the coefficient A, B, q are now function of the variable t and the last known coefficient of the complex root (in this case the $+$) has been replaced with the variable t.
By multiplying both member of the equation by the denominator of the left hand side we derive the following equation:

$$s^2+1 = \Big(A(t)s+B(t)\Big)(s-1) + q(t)\big(s^2-2s+t\big)$$

This expression allows now us to compute the parametric form of the function A, B, q; in particular considering that $s^2-2s+t=0$ and so $s^2 = 2s-t$ we can rewrite the previous expression as

$$2s-t+1 = A(t)(2s-t) - A(t)s + B(t)s - B(t) + \cancel{q(t)\big(2s-t-2s+t\big)}$$

$$\Rightarrow \quad \begin{cases} 2 & = A(t)+B(t) \\ -t+1 & = -tA(t)-B(t) \end{cases} \qquad \Rightarrow \quad A(t) = 1-\frac{2}{t-1}\,, \quad B(t) = 1+\frac{2}{t-1}$$

In order to have the complete the partial fraction expansion we need to compute the derivative (up to the second order) and evaluate them for $t=3$ (noting that $A(2)=-1$ and $B(2)=3$):

$$A' = \left.\frac{2}{(t-1)^2}\right|_{t=2} = 2 \qquad B' = -\left.\frac{2}{(t-1)^2}\right|_{t=2} = 2$$

$$A'' = -\left.\frac{4}{(t-1)^3}\right|_{t=2} = -4 \qquad B'' = \left.\frac{4}{(t-1)^3}\right|_{t=2} = 4$$

With all the coefficient defined is possible to express numerically the partial fraction expansion by using the equation 1.10:

$$\frac{2}{s-1} + \frac{-s+3}{(s^2-2s+2)^3} - 2\frac{s-1}{(s^2-2s+2)^2} + \frac{1}{2}\frac{-4s+4}{s^2-2s+2}$$

Example 1.7: computation of a partial fraction expansion

Considering the rational polynomial in s

$$\xi: \qquad \frac{P(s)}{Q(s)} = \frac{s(s^2+7)}{(s-1)(s+1)^3} = \frac{A}{s-1} + \frac{B}{s+1} + \frac{C}{(s+1)^2} + \frac{D}{(s+1)^3}$$

finding the partial fraction expansion means determining the coefficients $A, B, C, D \in \mathbb{R}$ that satisfies the equation. In order to solve this problem we can define ζ as the multiplication of ξ by $Q(s)$ and so

$$\zeta: \qquad s(s^2+7) = A(s+1)^3 + B(s-1)(s+1)^2 + C(s-1)(s+1) + D(s-1)$$

In order to find some coefficient we can note that such polynomial equation must be verified if we evaluate ζ for some *strategic* point, in particular for $s = 1$ (that's a root of Q) we have $8 = 2^3 A$ that determines $A = 1$, while if we evaluate for $s = -1$ we retrieve $-8 = -2D$ and so $D = 4$. We can so inflate ζ by moving on the left hand side the associated terms on A and D determining the rewritten expression:

$$\begin{aligned} \zeta_r: \quad s^3 + 7s - A(s^3 + 3s^2 + 3s + 1) - D(s-1) &= B(s-1)(s+1)^2 + C(s-1)(s+1) \\ s^3 + 7s - s^3 - 3s^2 - 3s - 1 - 4s + 4 &= B(s-1)(s^2 + 2s + 1) + C(s^2 - 1) \\ -3s^2 + 3 &= B(s^3 + s^2 - s - 1) + C(s^2 - 1) \end{aligned}$$

In order to determine the value of the other coefficient B, C we can start deriving ζ in the variable s:

$$\zeta' = \frac{d\zeta}{ds}: \qquad -6s = B(3s^2 + 2s - 1) + 2Cs$$

Evaluating this expression for $s = -1$ (associated to the root of both B and C) we have that $6 = -2Cs$ determining $C = -3$; after inflation we so have

$$\begin{aligned} \zeta_r': \qquad 0 &= B(3s^2 + 2s - 1) \\ \zeta'' = \frac{d\zeta'}{ds}: \qquad 0 &= B(6s + 2) \end{aligned}$$

With the last evaluation on $s = -1$ we obtain $0 = -4B$ requiring $B = 0$; with that said the overall partial fraction expansion is

$$\frac{s(s^2+7)}{(s-1)(s+1)^3} = \frac{1}{s-1} - \frac{3}{(s+1)^2} + \frac{4}{(s+1)^3}$$

1.5 Boundary value problem

Usually some differential equation problems requires some condition that are not defined at the initial time $t = 0$ of the time axes, but in a more general point $t = t^*$: in order to solve this problem a parametric approach is recomended. As example let's consider the following problem:

$$\begin{cases} x''(t) + y'(t) = \sin(t) \\ y'(t) + x'(t) = 1 + \sin(t) \\ x(0) = 0, \quad y(0) = -1, \quad x(1) = 1 \end{cases}$$

As we can see the last condition constrains the solution $x(t), y(t)$ at the time $t = 1$; in order to solve the problem in the Laplace domain we apply the transform on the system of differential equation and

so

$$\begin{cases} s^2\hat{x}(s) - x'(0) - sx(0) + s\hat{y}(s) - y(0) = \frac{1}{s^2+1} \\ s\hat{y}(s) - y(0) + s\hat{x}(s) - x(0) = \frac{1}{s} + \frac{1}{s^2+1} \end{cases}$$

If we now try to substitute the boundary condition on the problem we can see that we have no information for $x'(t)$ at the initial time: the idea now is to solve the problem parametrically (assign a variable, in this case A, to the initial condition $x'(0)$) and solve the problem this way. At the and we have to find the constant value A such that $x(1) = 1$. So by substituting we get

$$\begin{cases} s^2\hat{x}(s) - A + s\hat{y}(s) + 1 = \frac{1}{s^2+1} \\ s\hat{y}(s) + 1 + s\hat{x}(s) = \frac{1}{s} + \frac{1}{s^2+1} \end{cases}$$

To simplify the calculation of the algebraic solution of the functions $\hat{x}, \hat{y}$ we rewrite the system in matrix form that, by inversion, can determine the solutions (by using the Kramer method):

$$\begin{bmatrix} s^2 & s \\ s & s \end{bmatrix} \begin{pmatrix} \hat{x}(s) \\ \hat{y}(s) \end{pmatrix} = \begin{pmatrix} \frac{1}{s^2+1} + A - 1 \\ \frac{1}{s} + \frac{1}{s^2+1} - 1 \end{pmatrix}$$

$$\Rightarrow \qquad \hat{x}(s) = \frac{\det \begin{bmatrix} \frac{1}{s^2+1} + A - 1 & s \\ \frac{1}{s} + \frac{1}{s^2+1} - 1 & s \end{bmatrix}}{\det \begin{bmatrix} s^2 & s \\ s & s \end{bmatrix}} = \frac{sA - 1}{s^2(s-1)}$$

Having the boundary condition set on $x(t)$, our focus is to invert only the related transform $\hat{x}(s)$ (because, for the sake of the problem, it's not relevant determining $y(t)$); by doing the parametric partial fraction expansion we have

$$\frac{sA - 1}{s^2(s-1)} = \frac{c_1}{s-1} + \frac{c_2}{s} + \frac{c_3}{s^2}$$

where, by solving the linear system, the coefficients are equals to $c_1 = A - 1$, $c_2 = 1 - A$ and $c_3 = 1$. At this point we have everything in order to compute the inverted function of $\hat{x}(s)$:

$$x(t) = \mathscr{L}^{-1}\left\{\hat{x}(s)\right\}(t) = (A-1)e^t + (1-A) + t$$

The last thing to do now is to evaluate the function $x(t)$ at the time $t = 1$ and equal the result to 1 in order to determine the coefficient A related to the initial first derivative of x:

$$x(1) = (A-1)e + (1-A) + 1 = 1 \qquad \Rightarrow \quad A = 1$$

Example 1.8: boundary value problem

Considering the parametric differential equation problem depending on the parameter $A \in \mathbb{R}$

$$\begin{cases} x''(t) - x'(t) = t \\ x(0) = 1 \\ x'(0) = A \end{cases}$$

the solution can be found in the Laplace domain; starting off with the transformation of the problem using the $\mathscr{L}$ operator we have that

$$\mathscr{L}\left\{x''(t)\right\} = s^2\hat{x}(s) - x'(0) - sx(0) \qquad \mathscr{L}\left\{x'(t)\right\} = s\hat{x}(s) - x(0) \qquad \mathscr{L}\left\{t\right\} = \frac{1}{s^2}$$

$$\mathscr{L}\left\{x''(t) - x'(t) = t\right\} \quad \mapsto \quad s^2\hat{x}(s) - A - s - s\hat{x}(s) + 1 = \frac{1}{s^2}$$

Solving this expression for $\hat{x}(s)$ and performing the partial fraction expansion on the result (whose computation is here skipped) we obtain

$$\begin{aligned}\hat{x}(s) &= \frac{\frac{1}{s^2} + s + A - 1}{s^2 - s} = \frac{s^3 + As^2 - s^2 + 1}{s^3(s-1)} \\ &= \frac{A+1}{s-1} - \frac{A}{s} - \frac{1}{s^2} - \frac{1}{s^3}\end{aligned}$$

If we now want to compute the parameter A such that $x(1) = -\frac{1}{2}$ we need to invert $\hat{x}(s)$ in order to have the solution in the time domain. Considering so that

$$\mathscr{L}^{-1}\left\{\frac{1}{s-1}\right\} = e^t \qquad \mathscr{L}^{-1}\left\{\frac{1}{s}\right\} = 1 \qquad \mathscr{L}^{-1}\left\{\frac{1}{s^2}\right\} = t \qquad \mathscr{L}^{-1}\left\{\frac{1}{s^3}\right\} = \frac{t^2}{2!}$$

and so

$$\begin{aligned}x(t) &= \mathscr{L}^{-1}\left\{\hat{x}(s)\right\} = (A+1)e^t - A - t - \frac{t^2}{2} \\ x(1) &= (A+1)e - A - 1 - \frac{1}{2} = -\frac{1}{2}\end{aligned}$$

Solving the equation we retrieve the value of A that's -1.

Chapter 2

Constrained Minimization

Calculus revision

Given a function $f : \mathbb{R}^n \to \mathbb{R}$ it's possible to compute it's **minimum** by doing the assumption that f has a **Lipschitz continuos gradient**, notated as $f \in C^1(\mathbb{R}^n)$, meaning that

$$\exists\, \gamma > 0 \quad \text{such that} \quad \left\|\nabla f(\boldsymbol{x})^t - \nabla f(\boldsymbol{y})^t\right\| \leq \gamma \|\boldsymbol{x} - \boldsymbol{y}\| \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$$

A point $\boldsymbol{x}^* \in R$ is a **global minimum** if $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}$, file the point is a **local minimum** if $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x})$ for $\boldsymbol{x} \in B(\boldsymbol{x}^*, \delta)$. In particular the minimum is **strict** defined then it means $f(\boldsymbol{x}^*) < f(\boldsymbol{x})$.

Necessary conditions Necessary (but not sufficient) condition for a point $\boldsymbol{x}^* \in \mathbb{R}^n$ to be a local minimum is that

$$\nabla f(\boldsymbol{x}^*)^t = 0 \qquad \Rightarrow$$

This relation does not give any information on the point if it's a minimum, a maximum or a saddle point and so a second order (or higher) derivative analyses is required.

Assuming a function $f \in C^2(\mathbb{R}^n)$ (2 derivative continuos), if a point $\boldsymbol{x}^* \in R^n$ is a local minimum then $\nabla f(\boldsymbol{x}^*) = 0$ and $\nabla^2 f(\boldsymbol{x}^*)$ is semi positive definite and so

$$\boldsymbol{d}^t \nabla^2 f(\boldsymbol{x}^*) \boldsymbol{d} \geq 0 \qquad \forall d \in \mathbb{R}^n$$

where $\nabla^2 f(\boldsymbol{x}^*)$ is the hessian matrix of the function f. This condition (as the previous one) is necessary but not sufficient to determine if $\boldsymbol{x}^*$ is a minimum or a saddle point. If the hessian is **positive defined**, so $\nabla^2 f(\boldsymbol{x}^*) > 0$, then the condition is also necessary and $\boldsymbol{x}^*$ is a strict local minimum. In particular if the eigenvalues associated to $\nabla^2 f(\boldsymbol{x}^*)$ are all positive, then the matrix is positive defined.

2.1 Constrained minimization: Lagrange multipliers

The problem now is not to minimize a function $f \in C^2(\mathbb{R}^n)$ in all it's domain, but while considering a number m of constraints defined by equations $h_k \in C^2(\mathbb{R}^n)$, so solving a problem in the form:

$$\begin{aligned} \text{minimize:} \quad & f(\boldsymbol{x}) \\ \text{with constraints :} \quad & h_k(\boldsymbol{x}) = 0 \qquad k = 1, \ldots, m \end{aligned}$$

Lagrange multiplier The hard analytical problem of the constrained minimization can be solved using the **theorem of the Lagrange multiplier**. Let's consider a function f to be minimized with a constraints map $\boldsymbol{h}$ (such that $f, \boldsymbol{h} \in C^2(\mathbb{R}^n)$) and let $\boldsymbol{x}^*$ a local minimum of f and satisfies all the

constraints (and so $\boldsymbol{h}(\boldsymbol{x}^*) = \boldsymbol{0}$) then if $\nabla \boldsymbol{h}(\boldsymbol{x}^*)$ has maximum rank, then there exists m scalar λ_k such that

$$\nabla f(\boldsymbol{x}^*) - \sum_{k=1}^{m} \lambda_k \nabla h_k(\boldsymbol{x}^*) = 0 \tag{2.1}$$

This problem reduces now to a form on where we need to compute the eigenvalues λ_k of the **lagrangian** $\mathcal{L}$ defined as

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) := f(\boldsymbol{x}) - \sum_{k=1}^{m} \lambda_k h_k(\boldsymbol{x}) \tag{2.2}$$

In general the hardest part of the problem is determine all the points $\boldsymbol{x}^*$ that satisfies the Lagrange multiplier conditions because that implies to solve a non linear system of equations that usually is very hard to explicitly express. However the second part of the problem is way much easier: we need in fact to compute the kernel (null space) of $\nabla \boldsymbol{h}(\boldsymbol{x}^*)$ and, in order to have a local minimum, we have also to check that the matrix $\nabla^2_{\boldsymbol{x}}\big(f(\boldsymbol{x}^*) - \boldsymbol{\lambda}\, \boldsymbol{h}(\boldsymbol{x}^*)\big)$ is semi positive defined.

Using the lagrangian definition, the constrained minimization problem can be reduced to the following system of equations:

$$\begin{cases} \nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = \nabla_{\boldsymbol{x}} f(\boldsymbol{x}) - \lambda^t \nabla_{\boldsymbol{x}} \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0} \\ \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0} \end{cases} \tag{2.3}$$

All the points $\boldsymbol{x}^*$ that satisfies this system are candidates to be local maximum/minimum (in fact by computing the gradient and setting it to zero we are indeed searching for the stationary points of the lagrangian).
At this point to discriminate if the stationary point is maximum or minimum we have to use the second order conditions and in particular we must consider that the matrix $\nabla^2_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda})$ is positive defined in the kernel of the constraints map $\boldsymbol{h}(\boldsymbol{x})$, and so such that

$$\boldsymbol{z}^t \nabla^2_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda})\, \boldsymbol{z} > 0 \qquad \forall \boldsymbol{z} \in \ker\{\nabla \boldsymbol{h}(\boldsymbol{x}^*)\} \tag{2.4}$$

First and second order necessary condition To summarise the first order necessary condition for the point to be a local minimum is that the gradient ∇f of the function to minimize should be inside the linear space generated by the gradients of the constraints:

$$\nabla f(\boldsymbol{x}^*) \in \operatorname{span}\big\{\nabla h_1(\boldsymbol{x}^*), \dots, \nabla h_m(\boldsymbol{x}^*)\big\}$$

The second order necessary condition is that the matrix $\nabla^2_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda})$ is semi positive defined (and so it has to satisfy equation 2.4). In particular this condition is necessary when we consider an inequality of the type $\geq$, while the condition is sufficient when $\nabla^2_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) > 0$.

Example 2.1: constrained minimization problem

Let's consider the problem on where we want to minimize the function $f : \mathbb{R}^2 \to R$ using the constraint h defined as

$$f(x, y) = e^{x^2 - y^2} \qquad , \qquad h(x, y) = x - y^2$$

In order to solve this problem we at first need to build the lagrangian (having only one constraint $\boldsymbol{\lambda}$ reduces to a scalar) and so

$$\mathcal{L}(x, y, \lambda) = e^{x^2 - y^2} - \lambda\big(x - y^2\big)$$

We now need to compute the stationary points of the lagrangian and this means solving the following non linear system of equations:

$$\begin{cases} \nabla_x \mathcal{L}(x, y, \lambda) = 2x e^{x^2 - y^2} - \lambda = 0 \\ \nabla_y \mathcal{L}(x, y, \lambda) = -2y e^{x^2 - y^2} + 2\lambda y = 0 \\ \nabla_\lambda \mathcal{L}(x, y, \lambda) = -x + y^2 = 0 \end{cases}$$

$$\Rightarrow \quad (x,y,\lambda) \quad = \quad (0,0,0)\,, \left(\frac{1}{2}, \frac{1}{\sqrt{2}}, e^{-\frac{1}{4}}\right), \left(\frac{1}{2}, -\frac{1}{\sqrt{2}}, e^{-\frac{1}{4}}\right)$$

To determine now it this points are local maximum or minimum we have to firstly define the general gradient of the constraint map and then the hessian of the Lagrangian in respect to the variable $\boldsymbol{x} = (x, y)$:

$$\nabla h(x,y) = (1, -2y)$$
$$\nabla^2_{(x,y)} \mathcal{L} = \begin{bmatrix} (4x^2+2)e^{x^2-y^2} & -4xye^{x^2-y^2} \\ -4xye^{x^2-y^2} & (4y^2-2)e^{x^2-y^2} + 2\lambda \end{bmatrix}$$

Now we have to check each stationary point independently:

1. when $x = y = \lambda = 0$ we have that $\nabla h = (1, 0)$ while $\nabla^2_{(x,y)} \mathcal{L} = \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}$. By computing the null space of the vector $(1, 0)$ we can see that all the vectors in the form $(0, \alpha)$ match the definition; we can now check if the point is of maximum/minimum be determining if the matrix $\nabla^2_{(x,y)} \mathcal{L}$ is positive or negative defined:

$$\begin{pmatrix} 0 & \alpha \end{pmatrix} \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix} \begin{pmatrix} 0 \\ \alpha \end{pmatrix} = -2\alpha^2 \leq 0 \qquad \forall \alpha \in \mathbb{R}$$

   The hessian matrix is negative defined and so the point $(x, y) = (0, 0)$ is a local maximum.

2. evaluating for the second point $x = \frac{1}{2}$, $y = \frac{1}{\sqrt{2}}$ and $\lambda = e^{-\frac{1}{4}}$ we can compute the gradient $\nabla h = (1, -\sqrt{2})$ of the constraint map that determines a null space of the form $(\alpha\sqrt{2}, \alpha)$. Given the hessian matrix of the transform we can see that it's positive defined, in fact

$$e^{-\frac{1}{4}} \begin{pmatrix} \alpha\sqrt{2} & \alpha \end{pmatrix} \begin{bmatrix} 3 & -\sqrt{2} \\ -\sqrt{2} & 2 \end{bmatrix} \begin{pmatrix} \alpha\sqrt{2} \\ \alpha \end{pmatrix} = 4e^{-\frac{1}{2}}\alpha^2 > 0 \qquad \forall \alpha \in \mathbb{R}$$

   This means that the point is a local minimum.

3. considering instead the last point $x = \frac{1}{2}$, $y = -\frac{1}{\sqrt{2}}$ and $\lambda = e^{-\frac{1}{4}}$ we have a similar gradient $\delta h = (1, \sqrt{2})$ that determines a kernel in the form $(\alpha\sqrt{2}, -\alpha)$. Evaluating the hessian on the null space base we can see that the matrix is positive defined, in fact

$$e^{-\frac{1}{4}} \begin{pmatrix} \alpha\sqrt{2} & -\alpha \end{pmatrix} \begin{bmatrix} 3 & -\sqrt{2} \\ -\sqrt{2} & 2 \end{bmatrix} \begin{pmatrix} \alpha\sqrt{2} \\ -\alpha \end{pmatrix} = 4e^{-\frac{1}{2}}\alpha^2 > 0 \qquad \forall \alpha \in \mathbb{R}$$

We can see that the both points $\left(\frac{1}{2}, \frac{1}{\sqrt{2}}\right)$ and $\left(\frac{1}{2}, -\frac{1}{\sqrt{2}}\right)$ are local minimum and they are both also global minimum because we can see that $f\left(\frac{1}{2}, \frac{1}{\sqrt{2}}\right) = f\left(\frac{1}{2}, -\frac{1}{\sqrt{2}}\right) = e^{-\frac{1}{4}}$.

Example 2.2: determination of the non-linear system

Given the problem

$$\begin{aligned} \text{minimize:} \quad & f(x,y,z) = x - y + z^2 \\ \text{subject to:} \quad & h_1(x,y,z) = x^2 + y^2 - 2 = 0 \\ & h_2(x,y,z) = x + z - 1 = 0 \end{aligned}$$

the solution can be computed by firstly determining the lagrangian $\mathcal{L} = f - \lambda_1 h_1 - \lambda_2 h_2$ of the

problem that's

$$\mathcal{L}(x, y, z, \lambda_1, \lambda_2) = x - y + z^2 - \lambda_1(x^2 + y^2 - 2) - \lambda_2(x + z - 1)$$

The first order necessary condition related to the Lagrange multipliers states that a point, to be a minimum, must be a stationary one and so the candidate minimum points for this problem can be computed by solving the following system of non-linear equation:

$$\begin{cases} 1 - 2\lambda_1 x - \lambda_2 = 0 & : \frac{\partial \mathcal{L}}{\partial x} \\ -1 - 2\lambda_1 y = 0 & : \frac{\partial \mathcal{L}}{\partial y} \\ 2z - \lambda_2 = 0 & : \frac{\partial \mathcal{L}}{\partial z} \\ x^2 + y^2 = 2 & : \frac{\partial \mathcal{L}}{\partial \lambda_1} = h_1 \\ x + z = 1 & : \frac{\partial \mathcal{L}}{\partial \lambda_2} = h_2 \end{cases}$$

2.1.1 Sylvester theorem

The tedious and error prone operation of finding the minimum with the lagrangian multiplier is the one that's performed to determine if the matrix is (or is not) semi positive defined in the kernel $\ker\{\nabla \boldsymbol{h}(\boldsymbol{x}^*)\}$ of the gradient of the constraints map. In fact for every stationary point $\boldsymbol{x}^*$ of the lagrangian we have to check that

$$\boldsymbol{z}^t \, \nabla^2_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*) \, \boldsymbol{z} \geq 0 \qquad \forall \boldsymbol{z} \in \ker\{\nabla \boldsymbol{h}(\boldsymbol{x}^*)\}$$

For sake of simplicity from now we will denote the matrix $\nabla^2_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$ as A. We can note that the vector $\boldsymbol{z} \in \ker\{\nabla \boldsymbol{h}(\boldsymbol{x}^*)\}$ (and from now on we refer to the matrix $\nabla \boldsymbol{h}$ as B) can be expressed as a linear combination of the vectors $\boldsymbol{k}_i$ (that are representing the base of B) in the way

$$\boldsymbol{z} = \boldsymbol{k}_1 \alpha_1 + \boldsymbol{k}_2 \alpha_2 + \cdots + \boldsymbol{k}_p \alpha_p = K\boldsymbol{\alpha} \qquad \boldsymbol{\alpha} \in \mathbb{R}^p$$

We can see that this expression can be reduced to a multiplication of a matrix $K \in \mathbb{R}^{n \times p}$ (whose columns are the vector $\boldsymbol{k}_i$ of the kernel base) and a p-dimensional vector $\boldsymbol{\alpha}$ (where p is the number of constraints in the map $\boldsymbol{h}(\boldsymbol{x})$).

We can now rewrite the second order necessary condition considering that

$$\boldsymbol{z}^t A \boldsymbol{z} = \boldsymbol{\alpha}^t K^t A K \boldsymbol{\alpha} = \boldsymbol{\alpha}^t C \boldsymbol{\alpha} \qquad C := K^t A K \in \mathbb{R}^{p \times p}$$

Example 2.3

Let's consider the numerical problem when the values of the matrix $A = \nabla_{\boldsymbol{x}} \mathcal{L}$ and $B = \nabla \boldsymbol{h}$ are given with values

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 0 & 1 \\ 3 & 1 & 0 \end{bmatrix} \qquad B = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$$

The first thing now is to compute the manually compute the kernel of the matrix B in $\mathbb{R}^3$ solving the linear system

$$\begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \boldsymbol{0} \qquad \Rightarrow \qquad \begin{cases} z_1 = 0 \\ z_2 = \alpha \\ z_3 = \beta \end{cases} \qquad \alpha, \beta \in \mathbb{R}$$

At this point we can rewrite the kernel of B using the linear combination of the vector composing

the base:

$$\ker\{B\} = \begin{pmatrix} 0 \\ \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \alpha + \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \beta = \underbrace{\begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}}_{=K} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

The last thing is now to compute the matrix C that should be analyzed to know if it's (semi) positive defined or not:

$$K^t A K = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 2 & 0 & 1 \\ 3 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$C = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

This definition reduces the complexity of the problem at analyzing the matrix C (that's smaller than the original matrix A) and determining if that particular matrix is (semi) positive defined using two methods:

1. considering the fact that C is a symmetric matrix we know for sure that it can be diagonalised with an expression $T^t \Lambda T$ (where Λ is a diagonal matrix containing all the eigenvalues of C); considering the expression $\boldsymbol{\alpha}^t T^t \Lambda T \boldsymbol{\alpha}$ in order to have a semi positive defined matrix all the eigenvalue λ_i must be positive or at least equals to zero. If it happens that $\lambda_i > 0 \ \forall i$, then the matrix is positive defined and the point is a local minimum;
2. one other solution is to use the **Sylvester theorem** that states that *a symmetric matrix $A \in \mathbb{R}^{n \times n}$ is **positive defined** if and only if all the **principal minors** of A are strictly **positive*** (note that if one minor is equal to zero, no information can be retrieved with this method).

Example 2.4: application of the Sylvester theorem

Let's consider now the matrix

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 5 \end{bmatrix}$$

In order to determine if A is positive defined we have to compute all the principle minors starting from the first A_1 that's represented only by the element a_{11} of A and so

$$A_1 = \det\begin{bmatrix} 1 \end{bmatrix} = 1 > 0$$

The second (and last) principle minor of A is the determinant of the whole matrix and it happens that

$$A_2 = \det\begin{bmatrix} 1 & 2 \\ 2 & 5 \end{bmatrix} = 1 > 0$$

Having all minors greater then zero, this means that the matrix A is positive defined.

The same result can be achieved by determining all the eigenvalues of the matrix and showing they are all positive; we can in fact see that the characteristic polynomial of A is equal to

$$p(\lambda) = \det\left[A - \lambda I\right] = \det\begin{bmatrix} 1-\lambda & 2 \\ 2 & 5-\lambda \end{bmatrix} = \lambda^2 - 6\lambda + 1$$

$$\lambda_{1,2} = \frac{6 \pm \sqrt{36-4}}{2} = 3 \pm \frac{\sqrt{32}}{2}$$

nothing that $\frac{\sqrt{32}}{2} < \frac{\sqrt{36}}{2} = 3$, then all the eigenvalues λ_1, λ_2 are strictly positive and so A is positive defined.

Trick for semi positive matrices The Sylvester theorem gives no hint to determine if a matrix is semi-positive defined when a minor is equal to zero. However a way to determine if the matrix is semi-positive defined is by considering that the matrix $A + \varepsilon I$ should be positive defined for values of ε approaching zero from the positive direction (for $\varepsilon \to 0^+$).

Example 2.5: usage of the Sylvester theorem trick

Let's consider now the matrix

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$$

The first minor $A_1 = \det[1] = 1$ is positive, and so we have to compute the second principal minor noting that it's equal to zero:

$$A_2 = \det \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} = 4 - 2 \cdot 2 = 0$$

This relation gives no clue on determining if A is positive defined or not, but considering the trick yet defined we can compute

$$\det \begin{bmatrix} 1+\varepsilon & 2 \\ 2 & 4+\varepsilon \end{bmatrix} = (1+\varepsilon)(4+\varepsilon) - 4 = 5\varepsilon$$

We can see that this expression, for $\varepsilon \to 0^+$, is strictly greater then zero and this might conclude our analyses stating that A is semi positive defined.

The same result can be defined by computing the eigenvalues of A and so calculating the roots of the polynomial

$$p(\lambda) = \begin{bmatrix} 1-\lambda & 2 \\ 2 & 4-\lambda \end{bmatrix} = (1-\lambda)(4-\lambda) - 4 = \lambda^2 - 5\lambda \qquad \Rightarrow \quad \lambda_1 = 0, \lambda_2 = 5$$

Having one eigenvalues zero (and the other positive) determines that A is semi positive defined.

Example 2.6: counter example of the Sylvester theorem

Let's consider now the matrix

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

The first minor $A_1 = \det[1] = 1$ is positive, while the second one is zero, in fact

$$A_2 = \det \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = 0$$

This result gives no information but we can use the trick by considering the determinant

$$\det \begin{bmatrix} 1+\varepsilon & 1 \\ 1 & 1+\varepsilon \end{bmatrix} = (1+\varepsilon)^2 - 1 = \varepsilon^2 + 2\varepsilon > 0 \qquad \text{for } \varepsilon > 0$$

Being A a 3×3 matrix we also need to compute the third minor that leads to another zero:

$$A_3 = \det \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix} = 1 \det \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} - 1 \det \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} + 0 \det \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = 0$$

We can use another time the Sylvester theorem trick considering the determinant

$$p(\varepsilon) = \det \begin{bmatrix} 1+\varepsilon & 1 & 1 \\ 1 & 1+\varepsilon & 1 \\ 1 & 1 & \varepsilon \end{bmatrix} = (1+\varepsilon)\left(\varepsilon^2+\varepsilon-1\right) - (\varepsilon-1) - \varepsilon = \varepsilon\left(\varepsilon^2+2\varepsilon-2\right)$$

Having that the derivative $p'(\varepsilon) = 3\varepsilon^2 + 4\varepsilon - 2$ is negative for ε approaching zero, then we can conclude that A is not semi positive defined.

The same result can be confirmed calculating the eigenvalues of the matrix and so

$$p(\lambda) = \det \begin{bmatrix} 1-\lambda & 1 & 1 \\ 1 & 1-\lambda & 1 \\ 1 & 1 & -\lambda \end{bmatrix} = -\lambda\left(\lambda^2-2\lambda-2\right)$$

$$\Rightarrow \qquad \lambda_1 = 0 \qquad \lambda_{2,3} = \frac{2 \pm \sqrt{12}}{2} = 1 \pm \sqrt{3}$$

Noting that $\lambda_3 = 1 - \sqrt{3} < 0$ is negative, than the matrix A is for sure not semi positive defined.

2.2 Inequality constraints

Let's introduce now the problem of minimizing a function $f(\boldsymbol{x}) : \mathbb{R}^n \to R$ considering a set of p inequalities constrains in the form $g_k(\boldsymbol{x}) \geq 0$.

$$\begin{array}{ll} \text{minimize:} & f(\boldsymbol{x}) \\ \text{subject to:} & g_k(\boldsymbol{x}) \geq 0 \qquad k = 1, \ldots, p \end{array}$$

The first approach to this problem is by trying to transform the inequality constraints g_k into equality one (so having $h_k(\boldsymbol{x}) = 0$). This can be done considering that each constraint can be expressed as

$$g_k(\boldsymbol{x}) = \varepsilon_k^2 \geq 0 \qquad \Rightarrow h_k(\boldsymbol{x}, \varepsilon_k) = g_k(\boldsymbol{x}) - \varepsilon_k^2 = 0$$

Doing this process for each constraint we can determine a minimization problem that depends both on the function variable $\boldsymbol{x}$ but also on the so called **slack variables** $\boldsymbol{\varepsilon}$:

$$\begin{array}{ll} \text{minimize:} & f(\boldsymbol{x}) \\ \text{subject to:} & h_k(\boldsymbol{x}, \boldsymbol{\varepsilon}) = 0 \qquad k = 1, \ldots, p \end{array}$$

This kind of problem increases the computational costs because it increases the variable to minimize from n to $n + p$.

Note: We used the expression ε_k^2 and not ε_k because with this expression we don't need to specify one more inequality $\varepsilon_k \geq 0$.

Example 2.7

Let's consider the following problem:

$$\begin{array}{ll} \text{minimize:} & f(x,y) = x^2 \\ \text{subject to:} & x^2 + y^2 \leq 1 \\ & x + y \geq 0 \end{array}$$

In order to solve this problem we have to reduce the inequality constraints to equality ones; considering the first constraint, that has to be expressed in the form $g_1 \geq 0$ and so

$$g_1(x,y) = 1 - x^2 - y^2 \geq 0 \qquad \Rightarrow \quad h_1(x,y) = 1 - x^2 - y^2 - \varepsilon_1^2$$

At the same way it's possible to express the second inequality as the a constraint $h_2(x,y,\varepsilon_2) = x + y - \varepsilon_2^2$. With this expression being set the problem reduces to the form

$$\begin{aligned} \text{minimize:} \quad & f(x,y,\epsilon_1,\epsilon_2) = x^2 \\ \text{subject to:} \quad & h_1(x,y,\epsilon_1,\epsilon_2) = 1 - x^2 - y^2 - \epsilon_1^2 = 0 \\ & h_2(x,y,\epsilon_1,\epsilon_2) = x + y - \varepsilon_2^2 = 0 \end{aligned}$$

With the problem this stated we can build the lagrangian that depend's on both the original variables x, y but also the slack variables ϵ_1, ϵ_2 and so

$$\begin{aligned} \mathcal{L}(\underbrace{x,y,\epsilon_1,\epsilon_2}_{\boldsymbol{\epsilon}},\underbrace{\lambda_1,\lambda_2}_{\boldsymbol{\lambda}}) &= f(\boldsymbol{x}) - \lambda_1 h_1(\boldsymbol{x}) - \lambda_2 h_2(\boldsymbol{x}) \\ &= x^2 - \lambda_1\left(1 - x^2 - y^2 - \epsilon_1^2\right) - \lambda_2\left(x + y - \epsilon_2^2\right) \end{aligned}$$

To find the local minimum point we can start using the first necessary condition of the lagrangian multiplier, so by determining all the stationary point such that $\nabla\mathcal{L} = \mathbf{0}$; in practise this means solving the following non linear system of 6 equations in 6 variables:

$$\begin{cases} \frac{\partial\mathcal{L}}{\partial x} = & 2x - 2\lambda_1 x - \lambda_2 = 0 \\ \frac{\partial\mathcal{L}}{\partial y} = & 2\lambda_1 y - \lambda_2 = 0 \\ \frac{\partial\mathcal{L}}{\partial \epsilon_1} = & 2\lambda_1\epsilon_1 = 0 \\ \frac{\partial\mathcal{L}}{\partial \epsilon_2} = & 2\lambda_2\epsilon_2 = 0 \\ \frac{\partial\mathcal{L}}{\partial \lambda_1} = & x^2 + y^2 + \epsilon_1^2 - 1 = 0 \\ \frac{\partial\mathcal{L}}{\partial \lambda_2} = & x + y - \epsilon_2^2 = 0 \end{cases}$$

Solving minimization problems with inequality constraints can be tricky using the lagrangian multiplier method and so other algorithm (like the Karush-Kuhn-Tuckerthat's going to be explained) are usually preferred. However this kind of system can be solved by cases; considering in fact the second and third equations we can discriminate the solution considering the various combination when ϵ_1, ϵ_2 are equal (or not) to zero.
Considering for example the case $\epsilon_1 = \epsilon_2 = 0$ the non linear system reduces to the form

$$\begin{cases} 2x - 2\lambda_1 x - \lambda_2 = 0 \\ 2\lambda_1 y - \lambda_2 = 0 \\ x^2 + y^2 - 1 = 0 \\ x + y = 0 \end{cases}$$

Considering that the first two expressions are linear in respect to λ_1, λ_2 it's possible to determine their value directly in terms of x, y:

$$\lambda_1 = \frac{x}{y - x} \qquad \lambda_2 = 2\frac{xy}{y - x}$$

Considering instead the last 2 equations we have a non linear system that can be also computed by cases; from the last equation we have in fact that $y = -x$ and so

$$1 - x^2 - x^2 = 0 \qquad \Rightarrow \quad x = \pm\frac{1}{\sqrt{2}}$$

At this end of this process we defined 2 stationary points for the lagrangian:

$$(x, y, \epsilon_1, \epsilon_2, \lambda_1, \lambda_2) = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0, 0, -\frac{1}{2}, \frac{\sqrt{2}}{2}\right), \left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0, 0, \frac{1}{2}, -\frac{\sqrt{2}}{2}\right)$$

Solving symbolically this system of equation with a computer the lonely real solution of the problem (that determines a local minimum) is expressed in the form

$$(x, y, \epsilon_1, \epsilon_2, \lambda_1, \lambda_2) = \left(0, \sqrt{-\epsilon_1^2 + 1}, \epsilon_1, \sqrt[4]{1 - \epsilon_1^2}, 1, 1\right)$$

In order for the square roots to exists the value ϵ_1^2 should fit inside the range $[0, 1]$ and so the local minimum of the problem can be found on the vertical line

$$(x, y) = (0, k) \qquad k \in [0, 1]$$

2.3 Karush-Kuhn-Tucker conditions

The **Karush-Kuhn-Tucker** (KKT) **conditions** are a set of necessary and sufficient conditions (first and second order) that allows to describe if a point $\boldsymbol{x}^*$ is a minimum point of a minimization problem subjected to equality and inequality constraints, so for the problem

$$\begin{array}{lll} \text{minimize:} & f(\boldsymbol{x}) & \\ \text{subject to:} & h_k(\boldsymbol{x}) = 0 & k = 1, \dots, m \\ & g_k(\boldsymbol{x}) \geq 0 & k = 1, \dots, p \end{array}$$

where $f, h_k, g_k : \mathbb{R}^n \to \mathbb{R}$ are real evaluated function.

In order to describe this conditions it's important to define **set of the active constraints** $\mathcal{A}$ as

$$\mathcal{A}(\boldsymbol{x}^*) = \{k \mid g_k(\boldsymbol{x}^*) = 0\}$$

This means that given the constraints $g \in C(\mathbb{R}^n)$ it's defined as *active* if it happens that the point $\boldsymbol{x}^*$ is such that $g(\boldsymbol{x}^*) = 0$, so in practical way it's in the *border* between activating and not activating the constraints.

Qualified constraints Given $\boldsymbol{g} \in C^2(\mathbb{R}^p, \mathbb{R}^n)$ and $\boldsymbol{h} \in C(\mathbb{R}^m, \mathbb{R}^n)$ respectively the inequality and equality constraints map, then a point $\boldsymbol{x}^*$ is defined **qualified** if the gradients of the active constraints (and the equality ones) are linearly independent, and so

$$\{\nabla g_k(\boldsymbol{x}^*) \ : \ k \in \mathcal{A}(\boldsymbol{x}^*)\}\{\nabla h_1(\boldsymbol{x}^*), \dots, \nabla h_m(\boldsymbol{x}^*)\} \quad : \text{are linearly independent}$$

Example 2.8: qualification of a point

Considering the minimization problem started in example 2.2 (page 21) stated as

$$\begin{array}{ll} \text{minimize:} & f(x, y, z) = x - y + z^2 \\ \text{subject to:} & h_1(x, y, z) = x^2 + y^2 - 2 = 0 \\ & h_2(x, y, z) = x + z - 1 = 0 \end{array}$$

and considering the candidate point of minimum (resulting from the solution of the non-linear system)

$$x^* = \frac{\sqrt{3}}{2} - \frac{1}{2} \qquad y^* = \frac{1}{2} + \frac{\sqrt{3}}{2} \qquad z^* = \frac{3}{2} - \frac{\sqrt{3}}{2} \qquad \lambda_1^* = \frac{1}{2} - \frac{\sqrt{3}}{2} \qquad \lambda_2^* = 3 - \sqrt{3}$$

In order to check if the chosen point is qualified it's necessary to check if the gradients of the equality constraint and the active ones (that in this case are not present because there aren't inequality constraints) are linearly independent. In order to solve this problem firstly we need to build the matrix H represent the gradients of the constraints:

$$H(x,y,z) = \begin{bmatrix} \nabla h_1(x,y,z) \\ \nabla h_2(x,y,z) \end{bmatrix} = \begin{bmatrix} 2x & 2y & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

$$\Rightarrow \qquad H^* = H(x^*,y^*,z^*) = \begin{bmatrix} \sqrt{3}-1 & 1+\sqrt{3} & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

It's straightforward to see that the lines of the matrix H^* are linearly independent, considering in fact the rank of H^* by analysing the columns we see that's equal to 2, so equalizing the number of the constraints.

2.3.1 First order necessary condition

Lef $f \in C^1(\mathbb{R}^n)$ a function to minimize subjected to the inequality $\boldsymbol{g} \in C^1(\mathbb{R}^n, \mathbb{R}^p)$ and equality $\boldsymbol{h} \in C^1(\mathbb{R}^n, \mathbb{R}^m)$ constraint maps. If $\boldsymbol{x}^*$ satisfy constraint qualification then necessary condition for $\boldsymbol{x}^*$ to be a local minima is that that exists $m+p$ scalars λ_i, μ_j such that all the following conditions are satisfied:

$$\begin{aligned} \nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) &= \mathbf{0} \\ h_k(\boldsymbol{x}^*) &= 0 \qquad k = 0, \dots, m \\ g_k(\boldsymbol{x}^*) &\geq 0 \qquad k = 0, \dots, p \\ \mu_k^* g_k(\boldsymbol{x}^*) &= 0 \qquad k = 0, \dots, p \\ \mu_k^* &\geq 0 \qquad k = 0, \dots, p \end{aligned} \tag{2.5}$$

where the lagrangian is in this case defined as

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) - \sum_{k=1}^{p} \mu_k g_k(\boldsymbol{x}) - \sum_{k=1}^{m} \lambda_k h_k(\boldsymbol{x})$$

Example 2.9

Solving minimization problem with the KKT conditions is generally more easier; considering the case of the minimization problem in example 2.7 the lagrangian will depend only on 4 parameter (instead of 6 as in that example):

$$\mathcal{L}(x, y, \mu_1, \mu_2) = x^2 - \mu_1(1 - x^2 - y^2) - \mu_2(x + y)$$

Using the first order KKT condition the main non linear system to solve is in the form

$$\begin{cases} \left.\begin{aligned} 2x + 2\mu_1 x - \mu_2 &= 0 \\ 2\mu_1 y - \mu_2 &= 0 \end{aligned}\right\} & \nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) \\ \left.\begin{aligned} \mu_1(1 - x^2 - y^2) &= 0 \\ \mu_2(x + y) &= 0 \end{aligned}\right\} & \mu_k g_k(\boldsymbol{x}) = 0 \\ \mu_1, \mu_2 \geq 0 \end{cases}$$

This system of four non linear equations (the conditions $\mu_1, \mu_2 \geq 0$ are *easier* to consider) can be solved by cases:

- if $\mu_1 = \mu_2 = 0$ then the last two equalities are always verified, while the first two becomes

$$\begin{cases} 2x = 0 \\ 0 = 0 \end{cases}$$

this means that $x = 0$, while y is a free parameter that must satisfy both the constraints $x + y \geq 0$ (so that $y \geq 0$) and $1 - x^2 - y^2 \geq 0$ (and so $1 - y^2 \geq 0$ that determines $-1 \leq y \leq 1$): considering both condition at the same time we have that

$$\begin{cases} y \geq 0 \\ -1 \leq y \leq 1 \end{cases} \quad \Rightarrow \quad \text{solution: } x = \mu_1 = \mu_2 = 0 \quad y \in [0, 1]$$

- when considering $\mu_1 \neq 0$ and $\mu_2 = 0$ the system of equations become

$$\begin{cases} 2x + 2\mu_1 x = 0 \\ 2\mu_1 y = 0 \\ \mu_1(1 - x^2 - y^2) = 0 \end{cases}$$

Based on the assumption of the value μ_1, from the second equation we determine that y must be equal to 0 and so the third equation becomes $1 - x^2 = 0$ whose solutions are ± 1. Considering that the first equation can be factorised as $2x(1 + \mu_1) = 0$ (and $x \neq 0$), then in order to satisfy the relation μ_1 must be -1, but this imply that μ_1 is negative and it's an unacceptable solution, so for $\mu_1 \neq 0, \mu_2 = 0$ no solutions of the system can be found;

- considering instead $\mu_1 = 0$ and $\mu_2 \neq 0$ the system becomes

$$\begin{cases} 2x - \mu_2 = 0 \\ -\mu_2 = 0 \\ \mu_2(x + y) = 0 \end{cases}$$

The second equation states that μ_2 must be equal to 0 and so the solution of this problem is the same as the first case considered;

- considering both variables $\mu_1, \mu_2 \neq 0$ we have to solve the full systems; considering the last two equations we have that

$$\begin{cases} x + y = 0 \\ 1 - x^2 - y^2 = 0 \end{cases} \quad \Rightarrow \quad x = -y \quad \Rightarrow \quad 1 - 2x^2 = 0$$

and so it means that the two possible solution for the systems are $(x, y) = \left(\pm\frac{1}{\sqrt{2}}, \mp\frac{1}{\sqrt{2}}\right)$. Considering the first case where $x = 1/\sqrt{2}$ the first two equations become

$$\begin{cases} \frac{2}{\sqrt{2}}(1 + \mu_1) - \mu_2 = 0 \\ -\frac{2}{\sqrt{2}}\mu_1 - \mu_2 = 0 \end{cases} \quad \Rightarrow \quad \mu_2 = -\frac{2}{\sqrt{2}}\mu_1$$

$$\Rightarrow \quad \frac{4}{\sqrt{2}}\mu_1 + \frac{2}{\sqrt{2}} = 0 \quad \Rightarrow \quad \mu_1 = -\frac{1}{2} \qquad \mu_2 = \frac{1}{\sqrt{2}}$$

In this case $\mu_1 < 0$ is an unacceptable solution of the problem, and also considering the point $x = -1/\sqrt{2}$ the system becomes with no considerable solutions

$$\begin{cases} -\frac{2}{\sqrt{2}}(1 + \mu_1) - \mu_2 = 0 \\ \frac{2}{\sqrt{2}}\mu_1 - \mu_2 = 0 \end{cases} \quad \Rightarrow \quad \mu_1 = -\frac{1}{2} \quad \mu_2 = -\frac{1}{\sqrt{2}}$$

With all this considerations done the point of local minimum relies on the points

$$(x, y) = (0, k) \qquad k \in [0, 1]$$

2.3.2 Second order necessary conditions

Given a function $f \in C^1(\mathbb{R}^n)$ to minimize subject to the equality $\boldsymbol{h} \in C^1(\mathbb{R}^n, \mathbb{R}^m)$ and inequality $\boldsymbol{g} \in C^1(\mathbb{R}^n, \mathbb{R}^p)$ constraint maps, then necessary conditions for a point $\boldsymbol{x}^*$ (that satisfies the constraints) to be a local minima is that exists $m + p$ scalars λ_i, μ_i such that satisfy the first order conditions (equation 2.5) and

$$\boldsymbol{z}^t \nabla^2_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)\, \boldsymbol{z} \geq 0 \tag{2.6}$$

for all vector $\boldsymbol{z}$ such that

$$\begin{array}{lll} i) & \nabla h_k(\boldsymbol{x}^*)\boldsymbol{z} = 0 & k = 1, \dots, m \\ ii) & \nabla g_k(\boldsymbol{x}^*)\boldsymbol{z} = 0 & \text{for all } k \in \mathcal{A}(\boldsymbol{x}^*) \text{ and } \mu_k > 0 \\ iii) & \nabla g_k(\boldsymbol{x}^*)\boldsymbol{z} \geq 0 & \text{for all } k \in \mathcal{A}(\boldsymbol{x}^*) \text{ and } \mu_k = 0 \end{array}$$

In general the conditions $ii)$ and $iii)$ are hard to verify and so to have a less accurate necessary condition (by having a smaller set of possible vector $\boldsymbol{z}$), and so improving the chance of considering point that don't belong to the original domain, this two conditions can be substitute with the expression

$$\nabla g_k(\boldsymbol{x}^*)\boldsymbol{z} = 0 \qquad \text{for all } k \in \mathcal{A}(\boldsymbol{x}^*)$$

In other words the second order necessary condition (and similarly the sufficient one) states (from equation 2.6) that the the matrix $\nabla^2_{\boldsymbol{x}} \mathcal{L}$ must be semi-positive defined in the kernel of the active and equality constraints.

Example 2.10: kernel of the active constraints

Continuing examples 2.2 and 2.8 (page 27), on the minimum candidate point (x^*, y^*, z^*) we have the following matrix for the active constraints:

$$H^* = \begin{bmatrix} \sqrt{3}-1 & 1+\sqrt{3} & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

Computing the kernel of the active constraints means so solving the following system of linear equation determined by $H^*\boldsymbol{x} = \boldsymbol{0}$ and so

$$\begin{bmatrix} \sqrt{3}-1 & 1+\sqrt{3} & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

resulting in

$$\begin{cases} (\sqrt{3}-1)x + (1+\sqrt{3})y = 0 \\ x + z = 0 \end{cases}$$

Having 3 unknowns but only 2 equation means that we have to solve the problem in a parametric form; chosen $z = t$ we have $x = -t$ and so $y = t\frac{\sqrt{3}-1}{\sqrt{3}+1}$; choosing $t = \sqrt{3}+1$ (to simplify the results) we obtain the generator K of the kernel as

$$K = \begin{pmatrix} -\sqrt{3}-1 \\ \sqrt{3}-1 \\ \sqrt{3}+1 \end{pmatrix}$$

2.3.3 Second order sufficient conditions

Given a function $f \in C^1(\mathbb{R}^n)$ to minimize subject to the equality $\boldsymbol{h} \in C^1(\mathbb{R}^n, \mathbb{R}^m)$ and inequality $\boldsymbol{g} \in C^1(\mathbb{R}^n, \mathbb{R}^p)$ constraint maps, then sufficient condition for a point $\boldsymbol{x}^*$ (that satisfies the constraints) to

be a local minima is that exists $m+p$ scalars λ_i, μ_i such that satisfy the first order conditions (equation 2.5) and

$$\boldsymbol{z}^t \nabla^2_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)\, \boldsymbol{z} > 0 \tag{2.7}$$

for all vector $\boldsymbol{z}$ such that

$$\begin{array}{lll} i) & \nabla h_k(\boldsymbol{x}^*)\boldsymbol{z} = 0 & k = 1, \dots, m \\ ii) & \nabla g_k(\boldsymbol{x}^*)\boldsymbol{z} = 0 & \text{for all } k \in \mathcal{A}(\boldsymbol{x}^*) \text{ and } \mu_k > 0 \\ iii) & \nabla g_k(\boldsymbol{x}^*)\boldsymbol{z} \geq 0 & \text{for all } k \in \mathcal{A}(\boldsymbol{x}^*) \text{ and } \mu_k = 0 \end{array}$$

As in the previous case in order to have simpler calculation the third equation $iii)$ can be dropped (and so we obtain less accurate solution of the local minima).

Example 2.11: second order check

Given the problem started in example 2.2 (page 21) and carried on up to example 2.10 states as

$$\begin{array}{ll} \text{minimize:} & f(x,y,z) = x - y + z^2 \\ \text{subject to:} & h_1(x,y,z) = x^2 + y^2 - 2 = 0 \\ & h_2(x,y,z) = x + z - 1 = 0 \end{array}$$

we found a candidate for minimum point in

$$x^* = \frac{\sqrt{3}}{2} - \frac{1}{2} \qquad y^* = \frac{1}{2} + \frac{\sqrt{3}}{2} \qquad z^* = \frac{3}{2} - \frac{\sqrt{3}}{2} \qquad \lambda_1^* = \frac{1}{2} - \frac{\sqrt{3}}{2} \qquad \lambda_2^* = 3 - \sqrt{3}$$

that present the matrix H^* for the active constraints with related kernel K:

$$H^* = \begin{bmatrix} \sqrt{3}-1 & 1+\sqrt{3} & 0 \\ 1 & 0 & 1 \end{bmatrix} \qquad K = \begin{pmatrix} -\sqrt{3}-1 \\ \sqrt{3}-1 \\ \sqrt{3}+1 \end{pmatrix}$$

In order to check the second order necessary/sufficient condition we have to compute the hessian matrix $\nabla^2_{\boldsymbol{x}}\mathcal{L}$ of the lagrangian that's

$$L = \nabla^2_{\boldsymbol{x}}\mathcal{L} = \begin{bmatrix} -2\lambda_1 & 0 & 0 \\ 0 & -2\lambda_1 & 0 \\ 0 & 0 & 2 \end{bmatrix} \quad \Rightarrow \quad L^* = \begin{bmatrix} \sqrt{3}-1 & 0 & 0 \\ 0 & \sqrt{3}-1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

We have now to determine if this resulting matrix is (semi-)positive defined or not in the kernel of active constraints in order to determine if the point is a minimum or not; this mean performing the following matrix multiplication:

$$\begin{aligned} \alpha K^t L^* K \alpha &= \alpha \begin{pmatrix} -\sqrt{3}-1 & \sqrt{3}-1 & \sqrt{3}+1 \end{pmatrix} \begin{bmatrix} \sqrt{3}-1 & 0 & 0 \\ 0 & \sqrt{3}-1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} -\sqrt{3}-1 \\ \sqrt{3}-1 \\ \sqrt{3}+1 \end{pmatrix} \alpha \\ &= \alpha \begin{pmatrix} -\sqrt{3}-1 & \sqrt{3}-1 & \sqrt{3}+1 \end{pmatrix} \begin{pmatrix} -2 \\ 4-2\sqrt{3} \\ \sqrt{3}+1 \end{pmatrix} \alpha \\ &= \alpha\big(2\sqrt{3}+2+4\sqrt{3}-4-6+2\sqrt{3}+3+2\sqrt{3}+1\big)\alpha \\ &= \big(10\sqrt{3}-4\big)\alpha^2 \end{aligned}$$

Observing that for $\alpha \neq 0$ the product is always greater then zero we have that the hessian L^* is positive defined in the kernel of the active constraints, satisfying the second order sufficient condition that allow so to state that

$$\boldsymbol{x}^* = \left(\frac{\sqrt{3}}{2} - \frac{1}{2}, \frac{1}{2} + \frac{\sqrt{3}}{2}, \frac{3}{2} - \frac{\sqrt{3}}{2} \right) \qquad \text{is a minimum}$$

Example 2.12

Given the problem

$$\begin{aligned} &\text{minimize:} && f(x,y) = x^2 - xy \\ &\text{subject to:} && g_1(x,y) = 1 - x^2 - y^2 \geq 0 \\ & && g_2(x,y) = 1 - (x-1)^2 - y^2 \geq 0 \end{aligned}$$

To solve this kind of problem using the KKT conditions we firstly need to build the lagrangian of the problem that's equal to

$$\mathcal{L}(x,y,\mu_1,\mu_2) = x^2 - xy - \mu_1(1 - x^2 - y^2) - \mu_2[1 - (x-1)^2 - y^2]$$

At this point it's possible to construct the non linear system of equation whose solution are the stationary points candidate to be local minimum:

$$\begin{cases} \left.\begin{aligned} 2x - y + 2\mu_1 x + 2\mu_2(x-1) &= 0 \\ -x + 2\mu_1 y + 2\mu_2 y &= 0 \end{aligned}\right\} & \nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda}) \\ \left.\begin{aligned} \mu_1(1 - x^2 - y^2) &= 0 \\ \mu_2[1 - (x-1)^2 - y^2] &= 0 \end{aligned}\right\} & \mu_k g_k(\boldsymbol{x}) = 0 \\ \mu_1, \mu_2 \geq 0 \end{cases}$$

This system can be solved by cases of the value of μ_i and in particular

- when $\mu_1 = \mu_2 = 0$ the system reduces to the form

$$\begin{cases} 2x - y = 0 \\ -x = 0 \end{cases} \qquad \Rightarrow \quad x = y = 0$$

  This solution satisfy the constraints, in fact $g_1(0,0) = 1 \geq 0$ and $g_2(0,0) = 0 \geq 0$, so this point can be used to compute the second order conditions;

- considering all the other cases of μ_j the only other analytical solution to the system, found with `Mathematica`, determines the point

$$(x,y,\mu_1,\mu_2) = \left(\frac{1}{2}, \frac{\sqrt{3}}{2}, \frac{1}{\sqrt{3}} - \frac{1}{2}, \frac{1}{2} - \frac{\sqrt{3}}{6}\right)$$

In order to consider now the second order necessary/sufficient conditions we have firstly to evaluate the gradient $\nabla \boldsymbol{g}$ of the inequality constraint map (in respect to the variables x, y) resulting in the matrix

$$\nabla \boldsymbol{g} = \begin{bmatrix} -2x & -2y \\ -2(x-1) & -2y^2 \end{bmatrix}$$

For the second order conditions is also important to define the hessian matrix of the lagrangian respect to the variables x, y:

$$\nabla^2_{\boldsymbol{x}}\mathcal{L} = \begin{bmatrix} 2 + 2\mu 1 + 2\mu 2 & -1 \\ -1 & 2\mu_1 + 2\mu_2 \end{bmatrix}$$

We have now to check the conditions for the two local minima candidates:

- in the first case when $x = y = \mu_1 = \mu_2 = 0$ the set of the active constraints is just the second inequality g_2: in fact we have that $g_1(0,0) = 1 \neq 0$ while $g_2(0,0) = 0$. In respect to this vector we have to compute it's kernel (to determine the direction z on which verify the KKT condition) and in particular

$$\nabla g_2(0,0) = \begin{pmatrix} 0 \\ 2 \end{pmatrix} \qquad \Rightarrow \quad \ker\{\nabla g_2(0,0)\} = \alpha \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

Evaluating the expression $\boldsymbol{z}^t \nabla^2_{\boldsymbol{x}} \mathcal{L}\, \boldsymbol{z}$ for $\boldsymbol{z} \in \ker\{\nabla g_2\}$ we obtain that

$$\alpha \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \alpha = 2\alpha^2$$

Example 2.13: problem from an exam

Given the problem

$$\begin{aligned} \text{minimize:} \quad & f(x,y,z) = x - y + z^2 \\ \text{subject to:} \quad & h(x,y,z) = x^2 + y^2 - 2 = 0 \\ & g(x,y,z) = x^2 - 1 \geq 0 \end{aligned}$$

the solution can be obtained by using the KKT condition; in order to use them we firstly need to define the lagrangian $\mathcal{L} = f - \lambda h - \mu g$ of the problem that in this is of the form

$$\mathcal{L}(x,y,z,\lambda,\mu) = x - y + z^2 - \lambda\left(x^2 + y^2 - 2\right) - \mu\left(x^2 - 1\right)$$

The first order necessary condition allows to build the system of non-linear equation whose solution are candidates to be minimum point of f (respect to the constrains set), and so we have

$$\begin{cases} 1 - 2\lambda x - 2\mu x = 0 & : \frac{\partial \mathcal{L}}{\partial x} \\ -1 - 2\lambda y = 0 & : \frac{\partial \mathcal{L}}{\partial y} \\ -2z = 0 & : \frac{\partial \mathcal{L}}{\partial z} \\ x^2 + y^2 - 2 = 0 & : h \\ x^2 - 1 \geq 0 & : g \\ \mu(x^2 - 1) = 0 & : \mu g \\ \mu \geq 0 & : \mu \end{cases}$$

Verifying that the point $\boldsymbol{x}^* = (1,1,0)$ is a candidate minimum with $\lambda^* = -\frac{1}{2}$ and $\mu^* = 1$ (substituting the values in the system, all the relation are satisfied), the associated set of active constraints is represented by both h (in fact $1 + 1 - 2 = 0$) and g (indeed $1 - 1 = 0$) whose gradient can be regarded as

$$\nabla \mathcal{A} = \begin{bmatrix} \nabla h \\ \nabla g \end{bmatrix} = \begin{bmatrix} 2x & 2y & 0 \\ 2x & 0 & 0 \end{bmatrix} \qquad \Rightarrow \quad H = \nabla \mathcal{A}(\boldsymbol{x}^*) = \begin{bmatrix} 2 & 2 & 0 \\ 2 & 0 & 0 \end{bmatrix}$$

The constraints are so qualified because the gradients are linearly independent and related kernel, obtained by solving the equation $H\boldsymbol{x} = \boldsymbol{0}$, gives a generator of the space in the form of $\boldsymbol{k} = (0,0,1)$.
In order to determine if $\boldsymbol{x}^*$ is a minimum point using the second order necessary/sufficient KKT conditions it's mandatory to compute the hessian $\nabla^2_{\boldsymbol{x}} \mathcal{L}$ of the lagrangian:

$$\nabla^2_{\boldsymbol{x}} \mathcal{L} = \begin{bmatrix} -2(\lambda + \mu) & 0 & 0 \\ 0 & -2\lambda & 0 \\ 0 & 0 & -2 \end{bmatrix} \qquad \Rightarrow \quad L = \nabla^2_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^*) = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

We now need to check if the matrix L is (semi-)positive defined in the kernel of the active constraints, and so we have to compute

$$\begin{aligned}\alpha \boldsymbol{k}^t L \boldsymbol{k} \alpha &= \alpha \begin{pmatrix} 0 & 0 & 1 \end{pmatrix} \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \alpha \\ &= \alpha \begin{pmatrix} 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix} \alpha \\ &= 2\alpha^2\end{aligned}$$

Observing that this product is always greater then zero for $\alpha \neq 0$, then it means that L is positive defined in the kernel of the active constraints and so

$$\boldsymbol{x}^* = (1, 1, 0) \qquad \text{is a minimum point}$$

Example 2.14: problem from an exam

Given the problem

$$\begin{aligned} \text{minimize:} \quad & f(x,y,z) = x^2 - y + z \\ \text{subject to:} \quad & h(x,y,z) = y - x = 1 \\ & g_1(x,y,z) = x \geq 1 \\ & g_2(x,y,z) = z \geq 1 \end{aligned}$$

the associated lagrangian is

$$\mathcal{L}(x,y,z,\lambda,\mu_1,\mu_2) = x^2 - y + z - \lambda(y - x - 1) - \mu_1(x-1) - \mu_2(z-1)$$

The non-linear system associated to the the first order necessary KKT condition is so

$$\begin{cases} 2x + \lambda - \mu_1 = 0 & : \frac{\partial \mathcal{L}}{\partial x} \\ -1 - \lambda = 0 & : \frac{\partial \mathcal{L}}{\partial y} \\ 1 - \mu_2 = 0 & : \frac{\partial \mathcal{L}}{\partial z} \\ y - x - 1 = 0 & : h \\ x - 1 \geq 0 & : g_1 \\ z - 1 \geq 0 & : g_2 \\ \mu_1(x-1) = 0 & : \mu_1 g_1 \\ \mu_2(z-1) = 0 & : \mu_2 g_2 \\ \mu_1, \mu_2 \geq 0 \end{cases}$$

The lonely solution of this system is the one

$$x^* = 1 \qquad y^* = 2 \qquad z^* = 1 \qquad \lambda^* = -1 \qquad \mu_1^* = \mu_2^* = 1$$

For such point the set of active constraint $\mathcal{A}$ is determined by all the constraints $\{h, g_1, g_2\}$ (in fact $g_1(\boldsymbol{x}^*) = g_2(\boldsymbol{x}^*) = 0$) whose gradient is so

$$H = \begin{bmatrix} \nabla h \\ \nabla g_1 \\ \nabla g_2 \end{bmatrix} = \begin{bmatrix} -1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

This is a 3×3 matrix with linearly independent columns/rows and so the kernel obtained by solving $Hz = \mathbf{0}$ consist only in the null vector $z = \mathbf{0}$. Determined the hessian

$$L = \nabla^2_{\boldsymbol{x}} \mathcal{L} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

we have that $z^t L z$ is always zero in the kernel of the active constraints, meaning that the necessary condition is satisfied, but not the sufficient and so we are in the *grey zone* on which the point might (or not) be a local minima.

Chapter 3

Minimization of a Functional

A classical problem in minimization is the brachistrochrone, so the fastest ultimate curve (meaning determining the fastest trajectory to move from a point A to a point B considering that the only action is the one of the gravity).

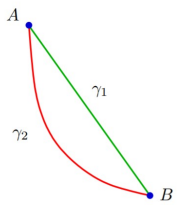


Figure 3.1: example of two trajectories γ_1, γ_2 that an object can follow for moving from point A to B.

Mathematically the problem is determining the curve $\mathcal{C} : [a, b] \to \mathbb{R}$ such that $\mathcal{C}(a) = y_a$ and $\mathcal{C}(b) = y_b$ that minimize the function $T(\mathcal{C})$ representing the time to travel and so

$$\text{minimize } T(\mathcal{C}) \qquad \text{for all possible curves } \mathcal{C}$$

We can see that a **function** takes a number as input and returns a number, like $f(x) = xe^x$. A **functional** is instead something that takes as input a function and returns a number and an example is

$$\mathcal{F}(x) = \int_a^b x(t)\, dt$$

where x is a function. Considering for example $x(t) = t^2$ the previous functional becomes

$$\mathcal{F}(x) = \int_a^b t^2\, dt = \frac{t^3}{3}\bigg|_a^b = \frac{b^3 - a^3}{3}$$

Another example of functional can be $\mathcal{G}(z) = z'(0) \int_0^1 z^2(t)\, dt$ and this expression can be evaluated for every generic function $z(t)$.

The question of this problem is how to define a *minimum* for a functional; in order to do so we have to firstly understand how to minimize a simple function and in particular given a function $f : A \subseteq \mathbb{R} \to \mathbb{R}$ has a minimum in the point x^* if

$$f(x^*) \leq f(x) \qquad \forall x \in A$$

Similarly given a function $\mathcal{F}(x)$ a *point* x^* (that in reality is a function) is a minimum if

$$\mathcal{F}(x^*) \leq \mathcal{F}(x) \qquad \forall x \in ?$$

In this case we have to specify the *class* (functional space) of functions we are considering, as example x can be a function that's continuous in the domain $[a, b]$ and so we can define

$$x \in C([a, b]) = \{g : [a, b] \to R \mid g \text{ is continuous}\}$$

In general changing the *domain* of the functional $\mathcal{F}$ may change the problem.

Example 3.1: domain change

Considering the function $f(x) = x^2 + 2$, it's roots can be computed if we allow complex solutions $z \in \mathbb{C}$ (and in fact $z = \pm i\sqrt{2}$), while if the domain of the solution is the real set $z \in \mathbb{R}$ no solutions exists.

Considering now the functional $\mathcal{F}$ defined as

$$\mathcal{F}(x) = \int_{-1}^{1} \left(x(t) - |t|\right)^2 dt$$

The $|t|$ introduce a cuspid in $t = 0$ that cannot be derived. If we minimize the functional in the continuous interval $x \in C([-1, 1])$, the solution is the function $x^*(t) = |t|$, in fact

$$\mathcal{F}(x^*) = \int_{-1}^{1} \left(|t| - |t|\right)^2 dt = \int_{-1}^{1} 0\, dt = 0$$

Choosing any other continuous function will result in a functional with a positive value.
Considering now to minimize the function in the domain of functions with continuous derivatives, and so $x \in C^1([-1, 1])$. In this case $|t| \notin C^1([-1, 1])$ (due to the cuspid). An approximation of the function $|t|$ that is continuous with also continuous first derivative is the function

$$x_\varepsilon(t) = \begin{cases} t & t \geq \varepsilon \\ \frac{1}{2}\left(\frac{t^2}{\varepsilon} + \varepsilon\right) & -\varepsilon < t < \varepsilon \\ -t & t \leq -\varepsilon \end{cases}$$

By pushing the limit $\varepsilon \to 0$ we can have the function that minimize the functional $\mathcal{F}$ in the domain $C^1([-1, 1])$. In fact we can see that the functional of x_ε becomes

$$\mathcal{F}(x_\varepsilon) = \int_{-\varepsilon}^{\varepsilon} \left(\frac{t^2}{2\varepsilon} + \frac{\varepsilon}{2} - |t|\right)^2 dt = \frac{\varepsilon^3}{10}$$

3.1 Analogies with linear algebra

To solve the problem of minimizing a functional we can see some relations with the linear algebra. The domain of the functional can be in fact see as a **functional space** $\mathbb{V}$ analogous to the vectorial one which the vectors are represented by the functions and the scalars are represented by real values. With this definition example of functional spaces might be

$$\mathbb{V}_1 = \{f : [0, 1] \to \mathbb{R} \text{ continuous}\}$$

$$\mathbb{V}_2 = \{f : [a, b] \to \mathbb{R} \text{ such that } f \in C^k([a, b])\} \qquad \text{with } a, b \in \mathbb{R},\ k \in \mathbb{N}$$

Given in fact two function $f, g \in \mathbb{V}$ that are member of the same functional space, we can see that each linear combination of the functions determine a function that's still in the vectorial space:

$$\alpha f(x) + \beta f(x) \in \mathbb{V} \qquad \forall \alpha, \beta \in \mathbb{R},\ f(x), g(x) \in \mathbb{V}$$

As example let's consider the two continuous function $f(x) = x^2$ and $g(x) = \sin x$, then we can clearly see that the function $2f(x) + \frac{1}{3}g(x) = 2x^2 + \frac{1}{3}\sin x$ is still continuous. This in general means that the functional space is *closed* respect to the operations of function summation and multiplication by a scalar.

Scalar product In linear algebra given two vector $\boldsymbol{v}_1, \boldsymbol{v}_2$ it exists the bilinear operator scalar product $\langle \boldsymbol{v}_1, \boldsymbol{v}_2 \rangle = \boldsymbol{v}_1 \cdot \boldsymbol{v}_2$ that satisfy the following rules:

$$\langle \alpha \boldsymbol{v} + \beta \boldsymbol{w}, \boldsymbol{z} \rangle = \langle \boldsymbol{z}, \alpha \boldsymbol{v} + \beta \boldsymbol{w} \rangle = \alpha \langle \boldsymbol{v}, \boldsymbol{z} \rangle + \beta \langle \boldsymbol{w}, \boldsymbol{z} \rangle \qquad \forall \alpha, \beta \in \mathbb{R},\ \boldsymbol{v}, \boldsymbol{w}, \boldsymbol{z} \in \mathbb{V}$$

$$\langle \boldsymbol{v}, \boldsymbol{v} \rangle \geq 0 \qquad \text{and} \qquad \langle \boldsymbol{v}, \boldsymbol{v} \rangle = 0 \quad \Leftrightarrow \quad \boldsymbol{v} = \boldsymbol{0}$$

Also in the functional space $\mathbb{V}$ can exists definitions of product scalar such the one here presented:

$$\langle f, g \rangle = \int_0^1 f(x) g(x)\, dx \tag{3.1}$$

Note: This is not the lonely function that can serve as product scalar of function and in this case the relation holds for the functional space $\mathbb{V} = \{f : [0,1] \to \mathbb{R} \text{ integrable}\}$.

In this case we can prove that this definitions meets the requirements stated for the scalar product considering the linear properties of the integrals as follows:

$$\begin{aligned} \langle \alpha f(x) + \beta g(x), h(x) \rangle &= \int_0^1 \Big(\alpha f(x) + \beta g(x) \Big) h(x)\, dx \\ &= \alpha \int_0^1 f(x) h(g)\, dx + \beta \int_0^2 g(x) h(x)\, dx \\ &= \alpha \langle f(x), h(x) \rangle + \beta \langle g(x), h(x) \rangle \end{aligned}$$

and also

$$\langle f(x), f(x) \rangle = \int_0^1 f(x) f(x)\, dx = \int_0^1 f^2(x) \geq 0$$

and in particular we can see that the scalar product $\langle f(x), f(x) \rangle$ will give as result o if and only if the function f is identically null, so such that $f(x) = 0$ for all x in it's domain (in this case $[0,1]$).

Norm In linear algebra it's also defined the norm of a vector as

$$\|\boldsymbol{v}\| := \sqrt{\boldsymbol{v} \cdot \boldsymbol{v}}$$

and this expression is used to compute, from a vector, a single positive value (and in particular $\|\boldsymbol{v}\| = 0$ if and only if $\boldsymbol{v} = \boldsymbol{0}$). This definition also holds for the functional space and considering the scalar product defined (as example) in equation 3.1 we can see that one definition of norm for function can be the one

$$\|f(x)\| = \sqrt{\langle f(x), f(x) \rangle} = \sqrt{\int_0^1 f^2(x)\, dx}$$

In this case we can clearly see that the norm $\|f(x)\| \geq 0$ is always positive defined (and is zero only in the case on which the function f is identically null).

Example 3.2: space of function

A space of functions can ve the one $f : [a,b] \to \mathbb{R}$ such that f is continuous, or for example $f : [a,b] \to \mathbb{R}$ where $f \in C^k([a,b])$ (for $k \in \mathbb{N}$). The same can be said for piecewise continuous functions.
Less trivially is a space of function the set of $f : [a,b] \to \mathbb{R}$ that are module integrable, and so all the function f such that

$$\int_a^b |f(x)|\, dx < \infty$$

In general we define as $L^p([a,b])$ the space of p-integrable functions, so such that

$$f \in L^p([a,b]) \qquad \Leftrightarrow \qquad \int_a^b |f(x)|^p\, dx < \infty$$

3.2 First variation

Given a functional $\mathcal{J} : \mathbb{V} \to \mathbb{R}$ defined in the function space $\mathbb{V}$, in order to determine the first necessary condition for minimum of the function we have to define the concept of **derivative** for functionals that is called (**first**) **variation** (or **directional derivative**).

To determine the first variation of the functional $\mathcal{J}$ respect to the function $x \in \mathbb{V}$ we have to define to compute the functional for the function $x + \alpha\eta$, where $\eta \in \mathbb{V}$ (and can be regarded as the ***direction of the derivative***) and $\alpha \in \mathbb{R}$. We can now denote the **first variation** as $\delta\mathcal{J}\big|_x : \mathbb{V} \to \mathbb{R}$ as the function that satisfy the following relation:

$$\mathcal{J}(x + \alpha\eta) = \mathcal{J}(y) + \delta\mathcal{J}\big|_x(\eta)\alpha + o(\alpha) \tag{3.2}$$

By using the definition of the small-o it's possible to express the limit relation that determines the first variation of the function as

$$\delta\mathcal{J}\big|_x(\eta) = \lim_{\alpha \to 0} \frac{\mathcal{J}(x + \alpha\eta) - \mathcal{J}(x)}{\alpha} \tag{3.3}$$

By so defining the function $g(\alpha) = \mathcal{J}(x + \alpha\eta)$ then it means that the first variation of the functional $\mathcal{J}$ respect the function x can be computed as

$$\delta\mathcal{J}\big|_x(\eta) = g'(0) \tag{3.4}$$

Example 3.3: directional derivative of a functional (first variation)

Given the functional

$$\mathcal{F}(x) = \int_0^1 \left(x^2(t) + 1\right) dt \ + x(1)$$

it's first variation respect to a generic direction $d(t)$ can be calculated by firstly determining the associated function $g : \mathbb{R} \to \mathbb{R}$ defined as

$$\begin{aligned} g(\alpha) &= \mathcal{F}\big(x(t) + \alpha\, d(t)\big) \\ &= \int_0^1 \left(\big(x(t) + \alpha\, d(t)\big)^2 + 1\right) dt + x(1) + \alpha\, d(1) \end{aligned}$$

As reported in equation 3.4, the first variation of the functional $\mathcal{F}$ can be regarded as the derivative of g respect to α evaluated for $\alpha = 0$; the first step is so determining

$$g'(\alpha) = \frac{d}{d\alpha} g(\alpha) = \frac{d}{d\alpha} \int_0^1 \left(\big(x(t) + \alpha\, d(t)\big)^2 + 1\right) dt + \frac{d}{d\alpha}\Big(x(1) + \alpha\, d(1)\Big)$$

Assuming that $\frac{d}{d\alpha}\int = \int \frac{d}{d\alpha}$ (operation that cannot always be performed) then the derivative of g can be regarded as

$$\begin{aligned} g'(\alpha) &= \int_0^1 \frac{d}{d\alpha}\Big(x(t) + \alpha\, d(t)\Big)^2 dt + \frac{d}{d\alpha}\Big(x(1) + \alpha\, d(1)\Big) \\ &= \int_0^1 2\big(x(t) + \alpha\, d(t)\big) d(t)\, dt + d(1) \end{aligned}$$

Evaluating this expression for $\alpha = 0$ then the first variation of $\mathcal{F}$ with direction $d(t)$ becomes

$$\delta\mathcal{F}\big|_x(d) = g'(0) = \int_0^1 2x(t)d(t)\, dt + d(1)$$

3.2.1 Optimality condition

Considering the minimization of a a function $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$, it has been shown that the first order necessary condition for a point $\boldsymbol{x}^* \in A$ to be a minimum point is that it's gradient $\nabla f(\boldsymbol{x}^*)$ (the

derivative) must be null. Similarly in calculus of variation it's proven that the **first order necessary condition** for the **optimality** of the solution is that the first variation of the functional $\mathcal{J}$ must be

$$\delta\mathcal{J}\big|_x(\eta) = 0 \tag{3.5}$$

for every *admissible perturbation* η.
In particular we define a **perturbation** $\eta \in \mathbb{V}$ **admissible** for the functional $\mathcal{J} \in A \subseteq \mathbb{V}$ respect to the function y^* if it happens that $y^* + \alpha\eta \in A$ for all value α *sufficiently close to* 0.

3.3 Fundamental lemma of the calculus of variations

Given a function $f : [a,b] \to \mathbb{R}$ (piecewise continuous) such that for all $g : [a,b] \to \mathbb{R}$ continuous (and more in particular $g \in C^\infty([a,b])$ to have a more general definition) with $g(a) = g(b) = 0$ and $g^{(k)}(a) = g^{(k)}(b) = 0\ \forall k$, if the integral

$$\int_a^b f(x)\, g(x)\, dx = 0 \tag{3.6}$$

then $f(x) = 0$ is identically null.

Lemma: sign permanence In order to later proof the fundamental lemma yet described, we have to remark the *sign permanence* that states: *given a function $f : [a,b] \to \mathbb{R}$ continuous such that $f(c) > 0$, then there exists a $\delta > 0$ such that*

$$f(x) \geq \frac{f(c)}{2} \qquad \forall x \in [c-\delta, c+\delta]$$

The proof of this lemma can be done defining the parameter $\epsilon = f(c)/2$; based on the assumption that $f(x)$ is continuous, then

$$\exists \delta > 0 \qquad \text{such that} \qquad |f(x) - f(c)| \leq \varepsilon = \frac{f(c)}{2} \qquad \forall |c - x| \leq \delta$$

Expanding the module operation the inequality becomes:

$$-\frac{f(c)}{2} = -\epsilon \leq f(x) - f(c) \leq \epsilon = \frac{f(c)}{2}$$

$$\Rightarrow \qquad \underbrace{-\frac{f(c)}{2} + f(c) \leq f(x)}_{\frac{f(c)}{2} \leq f(x)} \leq f(c) + \frac{f(c)}{2}$$

and so we prove the lemma of sign permanence.

Proof of the fundamental lemma The proof of the fundamental lemma of the calculus of variation can be done by contradiction. Let consider the integral $\int_a^b f(x)g(x)\,dx = 0$ for all functions $g(x)$ and such that, given the function f, exists a point c such that $f(c) > 0$ (in general the proof can be done for $f(c) \neq 0$). from the sign permanence lemma we can state that exists an interval $[c-\delta, c+\delta]$ such that $f(x) \geq \frac{f(c)}{2} \geq 0$. Determining the function g as

$$g(x) = \begin{cases} \dfrac{x - (c-\delta)}{\delta} & x \in [c-\delta, c] \\ \dfrac{c - x + \delta}{\delta} & x \in [c, c+\delta] \\ 0 & \text{otherwise} \end{cases}$$

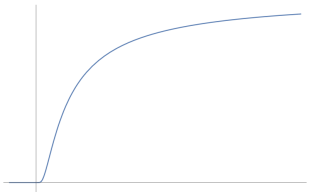


Figure 3.2: representation of $h(t)$ defined in equation 3.7

Note that this function is continuous but $g \notin C^\infty$ (later will be described a function that presents this feature). With this definition the integral $\int_a^b fg\,dx$ so becomes

$$\begin{aligned}\int_{c-\delta}^{c+\delta} f(x)g(x)\,dx &\geq \frac{f(c)}{2}\int_{c-\delta}^{c+\delta} g(x)\,dx \\ &\geq \frac{f(c)}{2}\frac{\delta}{2} > 0\end{aligned}$$

We can clearly see that the integral $\int_a^b fg\,dx$ has a positive real value grater than zero that's (considering that f has at least on point c such that $f(c) > 0$) in contraposition with the initial request that the integral should have been zero evaluated.

To create a direction function g that's in the set C^∞ we can use the function h(for whose those condition is demonstrated) as in figure 3.2 defined as:

$$h(t) = \begin{cases} e^{-1/t} & t > 0 \\ 0 & t \leq 0 \end{cases} \tag{3.7}$$

A way to create a function that presents a *bell shape* in the range $[0,1]$ is by computing $g(t) := h(t)h(1-t)$ how's graph in the range $[0,1]$ is similar to the one shown in figure 3.3.

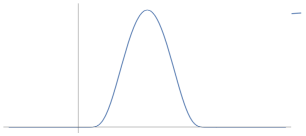


Figure 3.3: representation of the function $g(t) := h(t)h(1-t)$, where $h(t)$ is defined in equation 3.7

In particular to demonstrate the fundamental lemma of the calculus of variations we need to rescale the function g in order to have a bell centered in the point c with a *bell width* δ and so we consider

$$g(t) = K\,h\left(\frac{t-c+\delta}{2\delta}\right)h\left(\frac{\delta+c-t}{2\delta}\right)$$

where the constant $K \in \mathbb{R}$ is such that $g(c) = 1$ and $\int_a^b g(x)\,dx$. In this case we can see that $g \in C^\infty$, $g(x) \geq 0$ for all value x and specifically $g(x) = 0\,\forall x \notin [c-\delta, c+\delta]$.
The lemma can now be proven by contradiction; as in the previous case if we consider a function f that's not identically null (and in this case we assume that there is at least one point c on which f is positive), then we can state that (for the sign permanence theorem)

$$f(x) \geq \frac{f(c)}{2} \qquad \forall x \in [c-\delta, c+\delta]$$

We can now see that the original integral $\int_a^b fg\,dx$ is not equal to zero, in fact

$$\int_a^b f(x)g(x)\,dx \geq \frac{f(c)}{2}\int_a^b g(x)\,dx > 0$$

because $g(x)$ is always greater or equal to zero, determining a non zero value as result as in the previous case.

3.4 Euler Lagrange equation

Pendulum example The **Euler Lagrange equation** is the generalization of the minimum action principle that's used in physics. Considering as a practical example the motion of a pendulum of a mass m that's free to oscillate in respect to a pivot point using a rope of length l, the kinetic energy T and the potential term V of the system can be expressed as

$$T = \frac{1}{2}mv^2 \qquad\qquad V = mgy$$

Considering θ as the angle that the rope determines with the vertical axis, then the we can rewrite the energies as functions of the angular position θ and velocity $\dot{\theta}$ as

$$T(\theta, \dot{\theta}) = \frac{1}{2}ml^2\dot{\theta}^2 \qquad\qquad V(\theta) = -mgl\sin\theta$$

To solve the dynamic equation $\theta(t)$ of the mechanism we can compute the lagrangian L of the system defined as

$$L(\theta, \dot{\theta}) = T(\theta, \dot{\theta}) - V(\theta) = \frac{m}{2}l^2\dot{\theta}^2 + lmg\cos\theta$$

As law that's analyzed in mechanics physics we can state that the solution of the dynamics of the system is the one the function that minimize the **action** $\mathcal{A}$ of the system defined as

$$A(\theta) = \int_{t_0}^{t_1} L(\theta, \dot{\theta})\, dt \tag{3.8}$$

where the values $\theta(t_0) = \theta_0$ and $\theta(t_1) = \theta_1$ are known parameters.

In practice to determine the required solution we can use analytical tool to find the trajectory that can then be demonstrated to be the minimum of the action $\mathcal{A}$ (that's indeed a functional).

However we can also try to analytically determine the function $\theta^*(t)$ that minimize the functional $\mathcal{A}(\cdot)$ by computing the directional derivatives.

Let now consider the function $\theta^*(t)$ and a direction δ_θ that satisfy $\delta_\theta(t_0) = \delta_\theta(t_1) = 0$, we can then use the fundamental lemma of the calculus of variation to determine the minimal function (considering that the expression 3.8 of the action is comparable to equation 3.6 of the lemma).
To determine the minimum point we have in fact to determine the function θ^* whose first variation of the action $\dashv$ is zero for each direction of approach δ_θ to the point, and so in this example we need to compute

$$\begin{aligned}
\frac{d}{d\alpha}\mathcal{A}(\theta^* + \alpha\,\delta_\theta) &= \frac{d}{d\alpha}\int_{t_0}^{t_1} L(\theta^* + \alpha\,\delta_\theta, \dot{\theta}^* + \alpha\,\dot{\delta}_\theta) \\
&= \int_{t_0}^{t_1}\left[\frac{\partial}{\partial\theta}L(\dots)\frac{d}{d\alpha}(\theta^* + \alpha\,\delta_\theta) + \frac{\partial}{\partial\dot{\theta}}L(\dots)\frac{d}{d\alpha}(\dot{\theta}^* + \alpha\,\dot{\delta}_\theta)\right] dt \\
&= \int_{t_1}^{t_2}\left[(-lmg\sin(\theta^* + \alpha\,\delta_\theta))\delta_\theta + ml^2(\dot{\theta}^* + \alpha\,\dot{\delta}_\theta)\,\dot{\delta}_\theta\right] dt
\end{aligned}$$

Note that from in the first step we made the implicit assumption that $\frac{d}{d\cdot}\int = \int\frac{d}{d\cdot}$ while however this is not always possible. Evaluating the previous expression for $\alpha = 0$ gives the first variation of the action $\mathcal{A}$ that's

$$\delta\mathcal{A}\big|_{\theta^*}(\delta_\theta) = \frac{d}{d\alpha}\mathcal{A}(\theta^* + \alpha\,\delta_\theta)\Big|_{\alpha=0} = \int_{t_0}^{t_1}(-lmg\sin\theta^*\delta_\theta + ml^2\dot{\theta}^*\dot{\delta}_\theta)\,dt = 0 \qquad \forall\delta_\theta$$

Performing an integration by part allows to remove the term $\dot{\delta}_\theta$ that's hard to determine, and in fact

$$\frac{d}{dt}(ml^2\dot{\theta}^*\,\delta_\theta) = \frac{d}{dt}(ml^2\dot{\theta}^*)\,\delta_\theta + ml^2\dot{\theta}^*\dot{\delta}_\theta \quad\Rightarrow\quad ml^2\dot{\theta}^*\dot{\delta}_\theta = \frac{d}{dt}(ml^2\dot{\theta}^*\,\delta_\theta) - \frac{d}{dt}(ml^2\dot{\theta}^*)\,\delta_\theta$$

Performing the substitution of the integration by parts determines the result

$$\begin{aligned}\left.\frac{d\mathcal{A}}{d\alpha}\right|_{\alpha=0} &= \int_{t_0}^{t_1}\left[-lmg\sin\theta^* - \frac{d}{dt}\left(ml^2\theta^*\right)^2\right]\delta_\theta\,dt + \cancel{\left[ml^2\dot{\theta}^*\delta_\theta\right]\Big|_{t_0}^{t_1}} \\ &= \int_{t_0}^{t_1}\underbrace{\left[-lmg\sin\theta^* - \frac{d}{dt}\left(ml^2\dot{\theta}^*\right)\right]}_{f(t)}\delta_\theta\,dt = 0\end{aligned}$$

We can now see in this formulation that the marked expression represent the function f of the fundamental lemma considering that the relation must be true for all approaching direction δ_θ, and so it must be

$$-lmg\sin\theta^* - \frac{d}{dt}\left(ml^2\dot{\theta}^*\right) = 0$$

The problem now to complete the analyses of the pendulum motion is determining the function $\theta^*(t)$ that satisfy this expression and also match the boundary conditions $\theta(t_0) = \theta_0$ and $\theta(t_1) = \theta_1$.

3.4.1 General formulation

Given a functional

$$\mathcal{A}(x) = \int_a^b L\big(x(t), x'(t), t\big)\,dt \tag{3.9}$$

the problem is to minimize the functional $\mathcal{A}$ for a function $x \in \mathbb{V}$ subject to the boundary conditions $x(a) = x_a$ and $x(b) = x_b$ in the function space $\mathbb{V}$ defined as

$$\mathbb{V} = \left\{x \mid x \in C^2([a,b]) \text{ with } x(a) = x_a, x(b) = x_b\right\}$$

Let $\mathbb{D}$ the function space of all the feasible directions of derivatives defined as

$$\mathbb{D} = \left\{\delta_x \mid \delta_x \in C^\infty([a,b]) \text{ with } \delta_x(a) = \delta_x(b) = 0\right\}$$

the function $x(t)$ that minimize the functional is the one whose first variation is zero for any admissible perturbation, and so such that

$$\delta\mathcal{A}\big|_x(\delta_x) = \frac{d}{d\alpha}\mathcal{A}(x + \alpha\,\delta_x)\Big|_{\alpha=0} = 0 \qquad \forall\delta_x \in \mathbb{D}$$

Assuming the possibility to correctly apply the rule $\frac{d}{d\cdot}\int = \int\frac{d}{d\cdot}$ we can express the directional derivative of the functional $\mathcal{A}$ evaluated in $x + \alpha\,\delta_x$ as

$$\begin{aligned}\frac{d\mathcal{A}}{d\alpha} &= \frac{d}{d\alpha}\int_a^b L\big(x + \alpha\,\delta_x, x' + \alpha\,\delta_x', t\big)\,dt \\ &= \int_a^b\left(\frac{\partial L}{\partial x}\delta_x + \frac{\partial L}{\partial x'}\delta_x'\right)dt \\ \left.\frac{d\mathcal{A}}{d\alpha}\right|_{\alpha=0} &= \int_a^b\left(\frac{\partial L(x,x',t)}{\partial x}\delta_x + \frac{\partial L(x,x',t)}{\partial x'}\delta_x'\right)dt\end{aligned}$$

By performing the integration by part it's possible do reconvert the term associated do δ_x' into pieces depending on δ_x one of which, when evaluated, becomes zero due to the fact that $\delta_x(a) = \delta_x(b) = 0$:

$$\begin{aligned}\left.\frac{d\mathcal{A}}{d\alpha}\right|_{\alpha=0} &= \int_a^b\left[\frac{\partial L(x,x',t)}{\partial x} - \frac{d}{dt}\left(\frac{\partial L(x,x',t)}{\partial x'}\right)\right]\delta_x\,dt + \int_a^b\frac{d}{dt}\left(\frac{\partial L(x,x',t)}{\partial x'}\delta_x\right)dt \\ &= \int_a^b\left[\frac{\partial L(x,x',t)}{\partial x} - \frac{d}{dt}\left(\frac{\partial L(x,x',t)}{\partial x'}\right)\right]\delta_x\,dt + \cancel{\left[\frac{\partial L(x,x',t)}{\partial x'}\delta_x\right]\Big|_{\delta_x=a}^{b}}\end{aligned}$$

Note: In this case the substitution by part has been obtained by performing the operation

$$\frac{d}{dt}\left(\frac{\partial L(x,x',t)}{\partial x'}\delta_x\right) = \frac{d}{dt}\left(\frac{\partial L(x,x',t)}{\partial x'}\right)\delta_x + \frac{\partial L(x,x',t)}{\partial x'}\delta'_x$$

If we in fact explicit the term containing δ'_x we have

$$\frac{\partial L(x,x',t)}{\partial x'}\delta'_x = \underbrace{\frac{d}{dt}\left(\frac{\partial L(x,x',t)}{\partial x'}\delta_x\right)}_{\text{term \#1}} - \underbrace{\frac{d}{dt}\left(\frac{\partial L(x,x',t)}{\partial x'}\right)\delta_x}_{\text{term \#2}}$$

The second term still remains in the variation of the functional $\mathcal{A}$, while the first term is eliminated because it's evaluation becomes

$$\int_a^b \cancel{\frac{d}{dt}}\left(\frac{\partial L}{\partial x'}\delta_x\right)\cancel{dt} = \int_a^b \frac{\partial L}{\partial x'}d\delta_x = \frac{\partial L}{\partial x'}\Big(\delta_x(b) - \delta_x(a)\Big)$$

Looking ad the function space $\mathbb{D}$ we see that $\delta_x(a) = \delta_x(b) = 0$ and so the the evaluation of the integral becomes null.

We can now see that the condition to have a minimum for the functional $\mathcal{A}$, by using the fundamental lemma of calculus of variation, is requiring that the function f defined as follows is identically null:

$$\left.\frac{d\mathcal{A}(x+\alpha\,\delta_x)}{d\alpha}\right|_{\alpha=0} = \int_a^b \underbrace{\left(\frac{\partial L}{\partial x} - \frac{d}{dt}\frac{\partial L}{\partial x'}\right)}_{=f}\delta_x\,dt = 0 \tag{3.10}$$

This in general means that the function that minimise the functional $\mathcal{A}$ must solve the following second order ordinary differential equation:

$$\begin{cases} \dfrac{d}{dt}\dfrac{\partial L}{\partial x'} - \dfrac{\partial L}{\partial x} = 0 \\ x(a) = x_a\,,\ x(b) = x_b \end{cases}$$

Example 3.4: computation of the first variation

Given the generic functional defined as

$$\mathcal{F}(x) = \int_a^b G(x,x',t)\,dt + x(a)x(b)$$

in order to compute it's first variation we can use the *standard* approach by evaluating the perturbation of the functional respect to a function x:

$$\begin{aligned}
&= \left.\frac{d}{d\alpha}\mathcal{F}\big(x+\alpha\,\delta_x\big)\right|_{\alpha=0} \\
&= \frac{d}{d\alpha}\left(\int_a^b G\big(x+\alpha\,\delta_x, x'+\alpha\,\delta'_x, t\big)\,dt + \Big(x(a)+\alpha\delta_x(a)\Big)\Big(x(b)+\alpha\,\delta_x(b)\Big)\right) \\
&= \int_a^b \left(\frac{\partial G(\dots)}{\partial x}\,\delta_x + \frac{\partial G(\dots)}{\partial x'}\delta'_x\right)dt + \delta_x(a)\Big(x(b)+\alpha\,\delta_x(b)\Big) + \Big(x(a)+\alpha\,\delta_x(a)\Big)\delta_x(b) \\
&= \int_a^b \left(\frac{\partial G(\dots)}{\partial x}\,\delta_x + \frac{\partial G(\dots)}{\partial x'}\delta'_x\right)dt + \delta_x(a)x(b) + x(a)\delta_x(b)
\end{aligned}$$

This redundant formulation can be simplified by using the Gateaux derivative notation δ that

allows to express the first variation as

$$\begin{aligned}\delta\mathcal{F}(x) &= \delta\left(\int_a^b \mathcal{G}(x,x',t)\,dt + x(a)x(b)\right)\\ &= \int_a^b \delta G(x,x',t)\,dt + \delta\Big(x(a)x(b)\Big)\\ &= \int_a^b \left(\frac{\partial G(x,x',t)}{\partial x}\delta_x + \frac{\partial \mathcal{G}(x,x',t)}{\partial x'}\delta'_x\right)dt + \delta_{x(a)}x(b) + x(a)\delta_{x(b)}\end{aligned}$$

Extended definition

Considering now the functional $\mathcal{B}$ defined as

$$\mathcal{B}(x) = \int_a^b L(x(t),x'(t),x''(t),t)\,dt \tag{3.11}$$

the problem to solve now is the minimization of $\mathcal{B}(x)$ for all the function $x \in \mathbb{V}$ where the functional space is defined as

$$\mathbb{V} = \left\{ x \text{ such that } \begin{array}{l} x \in C^4([a,b]) \\ x(a) = x_a, x'(a) = x'_a, x(b) = x_b, x'(b) = x'_b \end{array} \right\}$$

and directional derivatives $\delta_x \in \mathbb{D}$ in the functional domain

$$\mathbb{D} = \left\{ \delta_x \text{ such that } \begin{array}{l} x \in C^\infty([a,b]) \\ x(a) = x'(a) = x(b) = x'(b) = 0 \end{array} \right\}$$

When trying to calculate the directional derivative of the functional $\mathcal{B}$ (skipping all the unnecessary computation that's similar to the cases yet studies) we end up to the following results:

$$\begin{aligned}&= \frac{d}{d\alpha}\bigg|_{\alpha=0} \mathcal{B}(x+\alpha,\delta_x)\\ &= \frac{d}{d\alpha}\bigg|_{\alpha=0} \int_a^b L\big(x+\alpha\,\delta_x, x'+\alpha\,\delta'_x, x''+\alpha\,\delta''_x, t\big)\,dt\\ &= \int_a^b \left(\frac{\partial L(x,x',x'',t)}{\partial x}\delta_x + \frac{\partial L(x,x',x'',t)}{\partial x'}\delta'_x + \frac{\partial L(x,x',x'',t)}{\partial x''}\delta''_x\right)dt\end{aligned}$$

To *cancel out* the terms involving the terms δ'_x, δ''_x it's necessary to use integration by parts considering the following derivatives:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial x'}\,\delta_x\right) = \frac{d}{dt}\frac{\partial L}{\partial x'}\delta_x + \frac{\partial L}{\partial x'}\delta'_x \qquad \frac{d}{dt}\left(\frac{\partial L}{\partial x''}\,\delta'_x\right) = \frac{d}{dt}\frac{\partial L}{\partial x''}\delta'_x + \frac{\partial L}{\partial x''}\delta''_x$$

and so with that said the derivative becomes

$$\begin{aligned}\delta\mathcal{B} &= \int_a^b \left(\frac{\partial L}{\partial x}\delta_x + \cancel{\frac{d}{dt}\left(\frac{\partial L}{\partial x'}\delta_x\right)} - \frac{d}{dt}\frac{\partial L}{\partial x'}\delta_x + \cancel{\frac{d}{dt}\left(\frac{\partial L}{\partial x''}\delta'_x\right)} - \frac{d}{dt}\frac{\partial L}{\partial x''}\delta'_x\right)dt\\ &= \int_a^b \left[\left(\frac{\partial L}{\partial x} - \frac{d}{dt}\frac{\partial L}{\partial x'}\right)\delta_x - \frac{d}{dt}\frac{\partial L}{\partial x''}\delta'_x\right]dt\end{aligned}$$

In the first line the terms are cancelled because if evaluated singularly we can see that they become in the form

$$\int_a^b \frac{d}{dt}\left(\frac{\partial L}{\partial x'}\delta_x\right)dt = \left[\frac{\partial L}{\partial x'}\delta_x\right]\Bigg|_a^b \xrightarrow{\delta_x(a)=\delta_x(b)=0} 0$$

To finish the analysis of the derivation we have to consider one last integration by part for the element

$$\frac{d}{dt}\left(\frac{d}{dt}\frac{\partial L}{\partial x''}\delta_x\right) = \frac{d^2}{dt^2}\frac{\partial L}{\partial x''}\delta_x + \frac{d}{dt}\frac{\partial L}{\partial x''}\delta'_x$$

$$\Rightarrow \qquad \delta\mathcal{B}(x) = \int_a^b \underbrace{\left(\frac{\partial L}{\partial x} - \frac{d}{dt}\frac{\partial L}{\partial x'} + \frac{d^2}{dt^2}\frac{\partial L}{\partial x''}\right)}_{=f}\delta_x\, dt \tag{3.12}$$

Using so the fundamental lemma of calculus of variation in order to determine the function x that minimize the functional $\mathcal{B}$ we need to solve the system of differential equation

$$\frac{\partial L}{\partial x} - \frac{d}{dt}\frac{\partial L}{\partial x'} + \frac{d^2}{dt^2}\frac{\partial L}{\partial x''} = 0$$

subjected to the the boundary conditions initially defined.

Example 3.5: minimization of a functional with the Euler-Lagrange method

Let's consider the problem

$$\begin{aligned} \text{minimize:} \quad & \int_a^b \left(\left(x''\right)^2 - x^2 + t\right) dt \\ \text{with:} \quad & x(a) = 0, x'(a) = 1 \quad x(b) = 1, x'(b) = 2 \end{aligned}$$

In this case the function to minimize is $L(x, x', x'', t) = (x'')^2 - x^2 + t$ and all the boundaries conditions are well defined. In this case to calculate the variation of the functional $\mathcal{A}(x) = \int_a^b L(x, x', x'', t)\, d$ by using equation 3.12:

$$\delta\mathcal{A} = \int_a^b \left(-2x - 0 + \frac{d^2x''}{dt^2}\right)\delta_x\, dt$$

Using the fundamental lemma of calculus of variation the terms in the bracket must be always equal to zero: this so represent, in conjunction with the boundary conditions, the ordinary differential equation that has to be solved to determine the solution of the problem:

$$\begin{cases} x^{(4)} - 2x = 0 \\ x(a) = 0, x'(a) = 1 \\ x(b) = 1, x'(b) = 2 \end{cases}$$

As we can see the differential equation is of order 4 and having 4 boundary condition and so it's possible to compute the solution.

3.4.2 Boundary conditions

Until now we considered minimization problems with all boundary conditions values set, while however this might not always be the case: equation 3.12 in fact is derived in the assumption of having all the variation constants $\delta_x = \delta_x^{(k)} = 0$ $(\forall k)$, but when this won't happens and so the general expression of the first variation is

$$\begin{aligned} \delta\mathcal{B} &= \int_a^b \left(\frac{\partial L}{\partial x} - \frac{d}{dt}\frac{\partial L}{\partial x'} + \frac{d^2}{dt^2}\frac{\partial L}{\partial x''}\right)\delta_x\, dt + \left[\left(\frac{\partial L}{\partial x'} - \frac{d}{dt}\frac{\partial L}{\partial x''}\right)\delta_x\right]_a^b + \left[\left(-\frac{\partial L}{\partial x''}\right)\delta'_x\right]_a^b \\ &= \int_a^b A\delta_x\, dt + B\delta_x(b) - C\delta_x(a) + D\delta'_x(b) - E\delta'_x(a) \end{aligned}$$

In this case if we consider that the boundary conditions fixed are only the one

$$x(a) = x_a \qquad x'(b) = x'_b$$

this also reflect on the functional domain of the variations $\delta_x \in \mathbb{D}$ that becomes

$$\tilde{\mathbb{D}} = \left\{\delta_x \in C^\infty([a,b]) \text{ such that } \delta_x(a) = \delta'_x(b) = 0\right\}$$

We can now see that in general the domain $\mathbb{D}$ of the variation set condition for value of the function (and it's derivative) on points only where the boundary conditions are defined. In the case described considering the values fixed it means that the solution for the minimum is the one that satisfies:

$$\delta\mathcal{B} = \int_a^b A\delta_x\, dt + B\delta_x(b) - E\delta'_x(a) \qquad \forall \delta_x \in \tilde{\mathbb{D}} \tag{3.13}$$

In this case we have 2 boundary degrees of freedom for the variation that we can define $\delta_x(b) = \delta_{xb}$ and $\delta'_x(a) = \delta'_{xa}$ (that are real evaluated variable); the simplest function that we might determine in order to create a variation $\delta_x(t)$ is a polynomial function depending on this parameters, and so

$$\delta_x(t) = \delta'_{xa}(t-a) + \frac{2\delta'_{xa}(a-b) + 3\delta_{xb}}{(a-b)^2}(t-a)^2 + \frac{\delta'_{xa}(a-b) + 2\delta_{xb}}{(a-b)^3}(t-a)^3$$

Considering that the Gateaux derivative $\delta\mathcal{B}$ should always be zero for each function $\delta_x(t)$ in it's domain, we can consider the special function with parameters $\delta_{xb} = 1$ and $\delta'_{xa} = 0$ becoming

$$\delta_x(t) = \frac{3}{(a-b)^2}(t-a)^2 + \frac{2}{(a-b)^3}(t-a)^3$$

Considering now that $\delta'_x(a) = 0$ we can see that in expression 3.13 the terms E is free and in order to have a null derivative it must be that

$$B = \left[\frac{\partial L}{\partial x'} - \frac{d}{dt}\frac{\partial L}{\partial x''}\right]\Bigg|_b = 0$$

Similarly we can create a particular polynomial (in particular considering $\delta_{xb} = \delta_x(b) = 0$ and $\delta'_{xa} = 1$) that set to zero the coefficients associated to B (and so leaving it *free* to change) and so, to have a zero evaluated derivative, it must be that

$$E = -\left.\frac{\partial L}{\partial x''}\right|_a = 0$$

In this case the differential equation associated to the terms B, E are called **trasversality conditions** and are necessary when the boundary conditions of the problem are not enough (in fact that expressions are evaluated at precise point on the *boarder* of the domain). At this point the solution of the original problem becomes

$$\begin{cases} \dfrac{\partial L}{\partial x} - \dfrac{d}{dt}\dfrac{\partial L}{\partial x'} + \dfrac{d^2}{dt^2}\dfrac{L}{\partial x''} = 0 \\ \left.\dfrac{\partial L}{\partial x'}\right|_b - \left.\dfrac{d}{dt}\dfrac{\partial L}{\partial x''}\right|_b = 0 \\ -\left.\dfrac{\partial L}{\partial x''}\right|_a = 0 \\ x(a) = x_a,\ x'(b) = x'_b \end{cases}$$

Example 3.6: minimization with less boundary conditions

Let's consider the problem of minimizing the functional $\mathcal{A}$ as in example 3.5 where the boundary

conditions in this case are only

$$x(a) = 0 \qquad\qquad x'(b) = 2$$

In this case the boundary conditions are not enough and the case is similar to the theory yet described: in this case the domain of the variations δ_x is described as

$$\mathbb{D} = \left\{\delta_x \in C^\infty([a,b]) \text{ such that } \delta_x(a) = \delta_x'(b) = 0\right\}$$

No information are set for the values $\delta_x(b), \delta_x'(a)$ that are so *free* to have different values in $\mathbb{R}$ so determining the following two trasverality conditions:

$$\left.\frac{\partial L}{\partial x'}\right|_b - \left.\frac{d}{dt}\frac{\partial L}{\partial x''}\right|_b = 0 - \left.2x'''\right|_b = -2x'''(b) = 0$$

$$-\left.\frac{\partial L}{\partial x''}\right|_a = -\left(-2x''\right)\Big|_a = 2x''(a) = 0$$

This, in conjunction with the differential equation determined by the function f in the integral and the known boundary conditions, determines the following ordinary system of equation that's the solution that minimize the functional:

$$\begin{cases} x^{(4)} - 2x = 0 \\ x'''(b) = 0 \\ x''(a) = 0 \\ x(a) = 0, x'(b) = 2 \end{cases}$$

Example 3.7: minimization of a functional

Let's consider the problem

$$\begin{aligned} \text{minimize:} \quad & \mathcal{A}(y) = \int_0^1 \left(\frac{\left(y'(x)\right)^2}{2} + y(x)y'(x) + y(x) \right) dx \\ \text{with:} \quad & y(1) = 1 \end{aligned}$$

In this case the lagrangian of the problem is defined as $L(y, y', x) = (y')^2/2 + yy' + y$; as formulated from page 43 the first variation of this functional so becomes

$$\delta\mathcal{A} = \int_0^1 \left(\frac{\partial L}{\partial y} - \frac{d}{dx}\frac{\partial L}{\partial y'} \right) \delta_y \, dx + \left[\frac{\partial L}{\partial y'} \delta_y \right] \Bigg|_{x=0}^{1} = 0$$

To formally compute the derivative we firstly need to define the domain of the variation δ_y that, having only one boundary condition, is

$$\mathbb{D} = \{\delta_y \in C^\infty([0,1]) \text{ with } \delta_y(1) = 0\}$$

At this point the Gateaux derivative can be computed as

$$\begin{aligned} \delta\mathcal{A} &= \int_0^1 \underbrace{\left(y' - (y'' - y')\right)}_{=f} \delta_y \, dx + \Big[(y' + y)\delta_y \Big]_0^1 \\ &= \int_0^1 -y''\delta_y \, dx - \underbrace{(y' + y)\delta_y(0)}_{\text{trasv. cond.}} \end{aligned}$$

As we can see we have that the f term (associated to y'') related to the fundamental lemma must be zero and so represent the differential equation associated to the solution of the problem while the second term $y' + y$ evaluated for $x = 0$ represent the trasverality condition that allow to have a unique solution of the system of ordinary differential equation that is:

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

By integration of the first differential equation we can determine the parametric solution of the system whose coefficients can be matched considering the boundary conditions:

$$y(x) = c_1 x + c_2$$

Substituting the parametric solution on the boundary conditions we can solve for the parameters:

$$\begin{cases} c_1 + 0\cancel{c_1} + c_2 = 0 \\ c_1 + c_2 = 1 \end{cases}$$

In this case the system of linear equation has no solution (in fact we have that $c_1 + c_2 = 0 \neq 1$) and so this means that the functional $\mathcal{A}$ cannot be minimized.

3.5 Functional minimization with constraints

Let's consider the problem of the minimization of a functional $\mathcal{F}$ subjected to an inequality constraints *at the boarder* as follows:

$$\begin{aligned} &\text{minimize:} && \mathcal{F}(z) = \int_a^b L(z, z', t)\, dt \\ &\text{subject to:} && w\big(z(a), z(b)\big) = 0 \end{aligned} \tag{3.14}$$

In this case we want to find the solutions for the function $z(t)$ in the functional spaced defined as

$$\mathbb{V} = \Big\{ z \in C^2([a,b]) \text{ such that } w\big(z(a), z(b)\big) = 0 \Big\}$$

In this case we cannot consider the linearity of the functional space (in fact two functions z_1, z_2 can satisfy the constraint b, but their sum $z_1 + z_2$ doesn't belong to the domain), and more difficult is the definition of the directional domain $\mathbb{D}$ on which we can compute all the possible derivatives.

Discretization approach A way to solve this problem is by discretizing the problem: given the domain $[a, b]$ we can subdivide him in n subintervals (in this case equally spaced) having length $h = \frac{b-a}{2}$; the axis t is so discretized in values

$$t_k = t_0 + kh = a + k\frac{b-a}{n}$$

With this definition we can compute the function z in the various point considering that $z(t_k) = z_k$ (and indeed we can also note that $t_0 = a$ and $t_n = b$). Using the mid-point squaring numerical method to integrate the original function, we can approximate the functional as

$$\mathcal{F}(z) = \int_a^b L(z, z', t)\, dt \approx h \sum_{k=1}^{n} L\left(z_{k-\frac{1}{2}}, z'_{k-\frac{1}{2}}, t_{k-\frac{1}{2}}\right)$$

where

$$z_{k-\frac{1}{2}} = \frac{z_k + z_{k-1}}{2} \qquad z'_{k-\frac{1}{2}} = \frac{z_k - z_{k-1}}{h} \qquad t_{k-\frac{1}{2}} = t_k - \frac{h}{2}$$

With this being said the initial constrained minimization problem (equation 3.14) can be discretized so obtaining the form

$$\begin{aligned} &\text{minimize:} && f(\boldsymbol{z}) = h\sum_{k=1}^{n} L\left(z_{k-\frac{1}{2}}, z'_{k-\frac{1}{2}}, t_{k-\frac{1}{2}}\right) \\ &\text{subject to:} && w\big(z_0, z_n\big) = 0 \end{aligned} \tag{3.15}$$

where $\boldsymbol{z} = \big(z_0, z_1, \ldots, z_n\big)^t$ is the vector off all the discretized values of the initial function $z(t)$. This formulation recall the constrained minimization problem with equality constraints (seen on page 19) and so we can use the Lagrange multiplier method by defining the expression

$$\mathcal{L}(\boldsymbol{z}, \lambda) = f(\boldsymbol{z}) - \lambda w(z_0, z_n)$$

To solve this kind of problem we need to find the stationary point of the yet built lagrangian $\mathcal{L}$,and so this means solving the following non-linear system determined by the equations

$$\frac{\partial \mathcal{L}}{\partial z_i} = 0 \qquad \forall i = 0, 1, \ldots, n \qquad\qquad \frac{\partial \mathcal{L}}{\partial \lambda} = 0$$

Starting with $i = 0$ we can compute that derivative of the lagrangian as

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial z_0} &= \frac{\partial}{\partial z_0}\left(h\sum_{k=1}^{n} L\left(\frac{z_k + z_{k-1}}{2}, \frac{z_k - z_{k-1}}{h}, t_{k-\frac{1}{2}}\right) - \lambda w(z_0, z_n)\right) \\ &= \frac{\partial}{\partial z_0}\left(hL\Big(\underbrace{\frac{z_1 + z_0}{2}, \frac{z_1 - z_0}{h}, t_{\frac{1}{2}}}_{=(1)}\Big) - \lambda w(z_0, z_n)\right) \\ &= h\frac{\partial L\big((1)\big)}{\partial z}\frac{\partial}{\partial z_0}\left(\frac{z_1 + z_0}{2}\right) + h\frac{\partial L\big((1)\big)}{\partial z'}\frac{\partial}{\partial z_0}\left(\frac{z_1 - z_0}{h}\right) - \lambda\frac{\partial w(z_0, z_n)}{\partial z_0} \\ &= \frac{h}{2}\frac{\partial L\big((1)\big)}{\partial z} - \frac{\partial L\big((1)\big)}{\partial z'} - \lambda\frac{\partial w(z_0, z_n)}{\partial z_0} \end{aligned}$$

Note that passing from the first to the second line only the terms associated to $k = 1$ present terms depending on z_0, and so only that part of the summation has been considered.

For all the other values $i \neq 0, n$, the mathematical expression of the derivative $\partial\mathcal{L}/\partial z_i$ becomes more *complex* (due to the fact that we have to consider two elements of the summation) and so we can use the simplified notation to express the partial terms such

$$\left.\frac{\partial \mathcal{L}}{\partial z}\right|_{k+\frac{1}{2}} := \frac{\partial L\left(\frac{z_{k+1}+z_k}{2}, \frac{z_{k+1}-z_k}{h}, t_{k+\frac{1}{2}}\right)}{\partial z}$$

With this, doing the steps as previously shown, we can compute the partial derivatives as

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial z_k} &= \frac{\partial}{\partial z_k}\left(h\sum_{j=1}^{n} L\left(\frac{z_j + z_{j-1}}{2}, \frac{z_j - z_{j-1}}{h}, t_{j-\frac{1}{2}}\right) - \lambda w(z_0, z_n)\right) \\ &= \frac{h}{2}\left(\left.\frac{\partial \mathcal{L}}{\partial z}\right|_{k-\frac{1}{2}} + \left.\frac{\partial \mathcal{L}}{\partial z}\right|_{k+\frac{1}{2}}\right) + \left.\frac{\partial \mathcal{L}}{\partial z'}\right|_{k-\frac{1}{2}} - \left.\frac{\partial \mathcal{L}}{\partial z'}\right|_{k+\frac{1}{2}} \\ &= h\left(\frac{\left.\frac{\partial \mathcal{L}}{\partial z}\right|_{k-\frac{1}{2}} + \left.\frac{\partial \mathcal{L}}{\partial z}\right|_{k+\frac{1}{2}}}{2} - \frac{\left.\frac{\partial \mathcal{L}}{\partial z'}\right|_{k+\frac{1}{2}} - \left.\frac{\partial \mathcal{L}}{\partial z'}\right|_{k-\frac{1}{2}}}{h}\right) \end{aligned}$$

The last partial derivative (computed for $k = n$) is instead

$$\frac{\partial \mathcal{L}}{\partial z_n} = \frac{h}{2} \left.\frac{\partial \mathcal{L}}{\partial z}\right|_{k-\frac{1}{2}} + \left.\frac{\partial \mathcal{L}}{\partial z'}\right|_{k-\frac{1}{2}} - \lambda \frac{\partial w(z_0, z_n)}{\partial z_n}$$

With all this calculation being done we determine that the non-linear system of equations representing the first order necessary condition for the minimum point of the lagrangian $\mathcal{L}$ is

$$\begin{cases} \dfrac{\left.\frac{\partial \mathcal{L}}{\partial z}\right|_{k-\frac{1}{2}} + \left.\frac{\partial \mathcal{L}}{\partial z}\right|_{k+\frac{1}{2}}}{2} - \dfrac{\left.\frac{\partial \mathcal{L}}{\partial z'}\right|_{k+\frac{1}{2}} - \left.\frac{\partial \mathcal{L}}{\partial z'}\right|_{k-\frac{1}{2}}}{h} = 0 & : A \\ \frac{h}{2} \left.\frac{\partial \mathcal{L}}{\partial z}\right|_{\frac{1}{2}} - \left.\frac{\partial \mathcal{L}}{\partial z'}\right|_{\frac{1}{2}} - \lambda \frac{\partial w(z_0, z_n)}{\partial z_0} = 0 & : B \\ \frac{h}{2} \left.\frac{\partial \mathcal{L}}{\partial z}\right|_{n-\frac{1}{2}} + \left.\frac{\partial \mathcal{L}}{\partial z'}\right|_{n-\frac{1}{2}} - \lambda \frac{\partial w(z_0, z_n)}{\partial z_n} = 0 & : C \end{cases}$$

By pushing the limit for $h \to 0$ (to have a *continuous discretization*), we can see a correlation of the two terms composing equation A, in fact

$$\frac{\left.\frac{\partial \mathcal{L}}{\partial z}\right|_{k-\frac{1}{2}} + \left.\frac{\partial \mathcal{L}}{\partial z}\right|_{k+\frac{1}{2}}}{2} \approx \frac{\partial \mathcal{L}}{\partial z}(z_k, z'_k, t_k) \qquad \frac{\left.\frac{\partial \mathcal{L}}{\partial z'}\right|_{k+\frac{1}{2}} - \left.\frac{\partial \mathcal{L}}{\partial z'}\right|_{k-\frac{1}{2}}}{h} \approx \frac{d}{dt}\left(\left.\frac{\partial \mathcal{L}}{\partial z'}\right|_k\right)$$

Similarly both the terms A and B presents the terms similar to the previous definition and they also presents another contribute due to the Lagrange multiplier λ. By so considering the limit $h \to 0$ we can see that $k = \frac{1}{2}$ tends to be a while $n - \frac{1}{2}$ tends to b, and so we can rewrite the systems of non-linear equations as

$$\begin{cases} \frac{\partial \mathcal{L}(z,z',t)}{\partial z} - \frac{d}{dt}\frac{\partial \mathcal{L}(z,z',t)}{\partial z'} = 0 \\ -\left.\frac{\partial \mathcal{L}(z,z',t)}{\partial z}\right|_a - \lambda \frac{\partial w(z(a),z(b))}{\partial z(a)} = 0 \\ -\left.\frac{\partial \mathcal{L}(z,z',t)}{\partial z}\right|_b - \lambda \frac{\partial w(z(a),z(b))}{\partial z(b)} = 0 \\ w\big(z(a), z(b)\big) = 0 \end{cases}$$

This is so the general formulation of the initial problem of minimizing a functional $\mathcal{F} = \int_a^b L\, dt$ subject to an equality constraint w.

Heuristic formulation

The same result can also be achieved with an heuristic formulation. Given so the problem of minimizing a functional $\mathcal{F}$ subjected to an equality constraint w (as in equation 3.14, page 49), the solution can be achieved by determining a new functional $\mathcal{L}$ defined as the lagrangian of the system:

$$\mathcal{L}(z, \lambda) = \int_a^b L(z, z', t)\, dt - \lambda w\big(z(a), z(b)\big) \tag{3.16}$$

At this point we can compute the variation of this functional considering it's Gateaux derivative δ

that's

$$\begin{aligned}
\delta\mathcal{L}(z,\lambda) = &\left.\frac{d}{d\alpha}\right|_{\alpha=0} \mathcal{L}\big(z+\alpha\,\delta_z, \lambda+\alpha\,\delta_\lambda\big) \\
=&\delta\int_a^b L(z,z',t)\,dt - \delta_\lambda\, w\big(z(a),z(b)\big) - \lambda\,\delta_w\big(z(a),z(b)\big) \\
=&\int_a^b \left(\frac{\partial\mathcal{L}}{\partial z}\delta_z + \frac{\partial\mathcal{L}}{\partial z'}\delta_z'\right) dt - \delta_\lambda\, w\big(z(a),z(b)\big) \\
&-\lambda\left(\frac{\partial w(z(a),z(b))}{\partial z(a)}\delta_{z(a)} + \frac{\partial w(z(a),z(b))}{\partial z(b)}\delta_{z(b)}\right) \\
=&\int_a^b \left(\frac{\partial\mathcal{L}}{\partial z} - \frac{d}{dt}\frac{\partial\mathcal{L}}{\partial z'}\right)\delta_z\,dt + \left[\frac{\partial\mathcal{L}}{\partial z'}\delta_z\right]\Bigg|_a^b - \delta_\lambda\, w\big(z(a),z(b)\big) \\
&-\lambda\left(\frac{\partial w}{\partial z(a)}\delta_{z(a)} + \frac{\partial w}{\partial z(b)}\delta_{z(b)}\right)
\end{aligned}$$

By evaluating the partial derivative $\partial\mathcal{L}/\partial z'$ and collecting common terms we can reduce the variation of the functional to the form

$$\delta\mathcal{L}(z,\lambda) = \int_a^b A\,\delta_z\,dt - \delta_\lambda\, w\big(z(a),z(b)\big) + B\delta_{z(a)} + C\delta_{z(b)}$$

where

$$A = \frac{\partial\mathcal{L}}{\partial z} - \frac{d}{dt}\frac{\partial\mathcal{L}}{\partial z'} \qquad B = -\left.\frac{\partial\mathcal{L}}{\partial z'}\right|_a - \lambda\frac{\partial w}{\partial z(a)} \qquad C = \left.\frac{\partial\mathcal{L}}{\partial z'}\right|_b - \lambda\frac{\partial w}{\partial z(b)}$$

Considering that the variation $\delta\mathcal{L}$ should be always be equal to zero $\forall\delta_z \in C^\infty([a,b]), \delta_\lambda \in \mathbb{R}$ we can consider that case on where $\delta_\lambda = \delta_{z(a)} = \delta_{z(b)} = \delta_z = 0$ for which, considering the fundamental lemma, we can state that the term A must be equal to zero. Similarly choosing $\delta_\lambda \neq 0$ and $\delta_{z(a)} = \delta_{z(b)} = 0$ the resultant condition becomes the initial equality constraint $w\big(z(a),z(b)\big) = 0$. Considering $\delta_\lambda = 0$ and $\delta_{z(a)} \neq 0$, $\delta_{z(b)} = 0$ we can state that B must be equal to zero and similarly, saying that $\delta_{z(b)} \neq 0$, that also C must be so. This means that the resultant system of non-linear equation that solves the problem can be expressed as

$$\begin{cases}
\dfrac{\partial\mathcal{L}}{\partial z} - \dfrac{d}{dt}\dfrac{\partial\mathcal{L}}{\partial z'} = 0 \\
w\big(z(a),z(b)\big) = 0 \\
\left.\dfrac{\partial\mathcal{L}}{\partial z'}\right|_a + \lambda\dfrac{\partial w}{\partial z(a)} = 0 \\
\left.\dfrac{\partial\mathcal{L}}{\partial z'}\right|_b - \lambda\dfrac{\partial w}{\partial z(b)} = 0
\end{cases} \tag{3.17}$$

We can observe that the resulting system is equivalent to the one obtained by pushing the limit $h \to 0$ with the discretized version previously performed.

Verification Considering the common case described by the problem

$$\begin{aligned}
&\text{minimize:} && \mathcal{F}(x) = \int_a^b L(x,x',t)\,dt \\
&\text{subject to:} && x(a) = x_a,\ x(b) = x_b
\end{aligned}$$

than the solution can be still achieved using the method yet shown. By in fact building the lagrangian

$$\mathcal{L}(x,\lambda_1,\lambda_2) = \int_a^b L(x,x',t)\,dt - \lambda_1\big(x(a)-x_a\big) - \lambda_2\big(x(b)-x_b\big)$$

substituting this function in the result of equation 3.17 we get the following differential system of equations:

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial x} - \frac{d}{dt}\frac{\partial \mathcal{L}}{\partial x'} = 0 \\ x(a) = x_a, \quad x(b) = x_b \\ -\lambda_1 - \left.\frac{\partial \mathcal{L}}{\partial x'}\right|_a = 0 \\ -\lambda_2 + \left.\frac{\partial \mathcal{L}}{\partial x'}\right|_b = 0 \end{cases}$$

The first two lines represent the terms that were always presents in the previous statement of the problem (without the equality constraints), while the last two depending from λ_i gives no real information because they can always be verified (in fact there will always exists a parameter λ_i that equals the derivative $\left.\frac{\partial \mathcal{L}}{\partial x'}\right|_{a,b}$); in general the variable λ_i can appear in the other equations, and so this last can be used to determine the stationary solution of the lagrangian $\mathcal{L}$.

Example 3.8: computation of first variation

Given the functional

$$\mathcal{F}(x) = x(0) + \int_0^1 tx^2 + \left(x'\right)^2 dt$$

it's first variation can be computed considering that $L(x, x', t) = tx^2 + \left(x'\right)^2$; equation 3.10 (page 44) helps determining the integral part by computing as function of the derivatives

$$\frac{\partial L}{\partial x} = 2tx \qquad \frac{d}{dt}\frac{\partial L}{\partial x'} = \frac{d}{dt}\left(2x'\right) = 2x''$$

To compute the complete variation of the system it's mandatory to consider the terms resulting from the integration by part (due to the fact that no boundary conditions are set) related to the term

$$\left[\frac{\partial L}{\partial x'}\delta_x\right]_a^b$$

With that said, the overall first variation can be computed as

$$\begin{aligned} \delta\mathcal{F} &= \delta_{x(0)} + 2\int_0^1 \left(tx - x''\right)\delta_x\, dt + 2x'\delta_{x(1)} - 2x'\delta_{x(0)} \\ &= 2\int_0^1 \left(tx - x''\right)\delta_x\, dt + 2x'\delta_{x(1)} + \left(1 - 2x'\right)\delta_{x(0)} \end{aligned}$$

Example 3.9: boundary value problem

Given the problem

$$\begin{aligned} &\text{minimize:} && \mathcal{F}(z) = x(0) + \int_0^1 x^2 + \left(x' - t\right)^2 dt \\ &\text{subject to:} && \int_0^1 x\, dt = 0 \\ & && x(1) = 2 \end{aligned}$$

the resulting boundary value problem can be computed considering the lagrangian $\mathcal{L}$ of the system that's

$$\begin{aligned}\mathcal{L}(x,\lambda,\mu) &= \mathcal{F}(x) - \lambda\int_0^1 x\,dt - \mu\Big(x(1)-2\Big)\\ &= \int_0^1 \underbrace{x^2 + \big(x'-t\big)^2 - \lambda x}_{L}\,dt + x(0) - \mu\big(x(1)-2\big)\end{aligned}$$

Starting from this point we can so compute the first variation of $\mathcal{L}$ that's:

$$\begin{aligned}\delta\mathcal{L} &= \int_0^1 \left(\frac{\partial L}{\partial x} - \frac{d}{dt}\frac{\partial L}{\partial x'}\right)\delta_x\,dt + \left[\frac{\partial L}{\partial x'}\delta_x\right]_0^1 - \int_0^1 x\,\delta_\lambda\,dt\\ &\quad + \delta_{x(0)} - \mu\delta_{x(1)} - \big(x(1)-2\big)\delta_\mu\\ &= \int_0^1 \Big(-2x'' + 2x + 2 - \lambda\Big)\delta_x\,dt + 2\big(x'(1)-1\big)\delta_{x(1)} - 2x'(0)\delta_{x(0)}\\ &\quad - \int_0^1 x\delta_\lambda\,dt + \delta_{x(0)} - \mu\delta_{x(1)} - \big(x(1)-2\big)\delta_\mu\end{aligned}$$

The resulting boundary valued problem is so determined by setting to zero the terms that multiply every variation $\delta_{\cdot}$ and so:

$$\begin{cases}\delta_x: & 2x'' - 2x - 2 + \lambda = 0\\ \delta_{x(0)}: & 2x'(0) = 1\\ \delta_{x(1)}: & \cancel{2x'(1) - 2 = \mu} \qquad \text{: trivially satisfied}\\ \delta_\lambda: & \int_0^1 x = 0\\ \delta_\mu: & x(1) = 2\end{cases}$$

Example 3.10: exam's simulation

Given the functional

$$\begin{aligned}&\text{minimize:} && \mathcal{F}(x,y,z) = y(1) + \int_0^1 x^2 + z^2 + x'y' + z'^2\,dt\\ &\text{subject to:} && x(1) = 2\end{aligned}$$

the resulting boundary value problem can be stated by determining the first variation of the lagrangian $\mathcal{L}$

$$\begin{aligned}
\mathcal{L}(x,y,z,\lambda) &= y(1) + \int_0^1 \underbrace{x^2 + z^2 + x'y' + z'^2}_{=L(x,y,z)}\, dt - \lambda\big(x(1)-2\big) \\
\delta\mathcal{L} &= \delta_{y(1)} + \int_0^1 \left(\frac{\partial L}{\partial x} - \frac{d}{dt}\frac{\partial L}{\partial x'}\right)\delta_x\, dt + \left[\frac{\partial L}{\partial x'}\delta_x\right]_0^1 \\
&\quad + \int_0^1 \left(\frac{\partial L}{\partial y} - \frac{d}{dt}\frac{\partial L}{\partial y'}\right)\delta_y\, dt + \left[\frac{\partial L}{\partial y'}\delta_y\right]_0^1 \\
&\quad + \int_0^1 \left(\frac{\partial L}{\partial z} - \frac{d}{dt}\frac{\partial L}{\partial z'}\right)\delta_z\, dt + \left[\frac{\partial L}{\partial z'}\delta_z\right]_0^1 + \lambda\delta_{x(1)} + \big(x(1)-2\big)\delta_\lambda \\
&= \delta_{y(1)} + \int_0^1 \big(2x - y''\big)\delta_x\, dt + y'(1)\delta_{x(1)} - y'(0)\delta_{x(0)} \\
&\quad + \int_0^1 -x''\delta_y\, dt + x'(1)\delta_{y(1)} - x'(0)\delta_{y(0)} \\
&\quad + \int_0^1 \big(2z - 2z''\big)\delta_z\, dt + 2z'(1)\delta_{z(1)} - 2z'(0)\delta_{z(0)} + \lambda\delta_{x(1)} + \big(x(1)-2\big)\delta_\lambda
\end{aligned}$$

Setting to zero the terms associated to each variation $\delta_\cdot$ determines the following boundary value problem, solution of the functional minimization:

$$\begin{cases}
2x - y'' = 0 & : \delta_x \\
x'' = 0 & : \delta_y \\
z - z'' = 0 & : \delta_z \\
y'(0) = 0 & : \delta_{x(0)} \\
\cancel{y'(1) + \lambda = 0} & : \delta_{x(1)}\text{, trivially solved in } \lambda \\
x'(0) = 0 & : \delta_{y(0)} \\
1 + x'(1) = 0 & : \delta_{y(1)} \\
z'(0) = 0 & : \delta_{z(0)} \\
z'(1) = 0 & : \delta_{z(1)} \\
x(1) = 2 & : \delta_\lambda
\end{cases}$$

3.6 The brachistochrone problem

As saw at the first part of this chapter (page 36), functional minimization can be used to solve real problem. As example the brachistochrone problem wants to determine the path that minimize the time T to reach point B from a point A for a mass m subjected only to the force of gravity: this means minimizing the function

$$T = \int_A^B \frac{1}{v}\, ds$$

where v is the velocity of the point over the arc length ds. In order to describe the problem we consider x as the horizontal axis and y the vertical one (with positive direction the one *facing upwards*); the point A is described by the coordinates $(x_a, x_b) = (0,0)$ (so it's in the center of the reference system) while the point $B = (x_b, y_b)$ is placed anywhere in the lower semi-plane (in fact we must require that $y_b < 0$ in order to have a free fall of the object).

Considering that at the initial position A both the kinetic and potential energy are zero, due to the

conservation of energy we can derive that

$$0 = mgy + \frac{1}{2}mv^2 \qquad \Rightarrow \quad v = \sqrt{-2gy}$$

We so have the velocity v as depending from the vertical coordinate y of the point, and so the next step is to relate the arc length ds with the horizontal variation dx: we can see that

$$ds = \sqrt{dx^2 + dy^2} = \sqrt{dx^2 + y'^2(x)dx^2} = \sqrt{1 + y'^2}dx$$

The initial brachistochrone problem can so be states as the following functional minimization problem:

$$\begin{aligned} &\text{minimize:} && \mathcal{F}(y, y', x) = \int_{x_a=0}^{x_b} \sqrt{\frac{1+y'^2}{-2gy}}\, dx \\ &\text{subject to:} && y(x_a) = y_a = 0 \qquad y(x_b) = y_b < 0 \end{aligned}$$

where the independent variable is the x coordinate. Defined as $L(y, y', x) = \sqrt{\frac{1+y'^2}{-2gy}}$ the integral part of the target function to be minimized, we can convert the functional minimization to the following boundary value problem:

$$\begin{cases} \dfrac{\partial L}{\partial y} - \dfrac{d}{dx}\dfrac{\partial L}{\partial y'} = -\dfrac{1 + 3y'^2 + 2y'^4 - 2y\,y''}{2\sqrt{2}g^2y^3\sqrt{\left(-\frac{1+y'^2}{gy}\right)^3}} = 0 \\ \\ y(0) = 0 \\ y(x_b) = y_b \end{cases}$$

Even if it's not clear at first sight, it can be proven that the solution of the differential equation is a cycloid described by the equations

$$x(\theta) = \frac{1}{2}k^2(\theta - \sin\theta) \qquad\qquad y(\theta) = \frac{1}{2}k^2(1 - \cos\theta)$$

where the constant k and the range of the new angle θ of the parametrization can be found by using the boundary conditions; in particular the initial condition on points A allows to state that $\theta_a = 0$ (in fact $x(0) = \frac{1}{2}k^2 0 = 0$ and also $y(0) = 0$), while considering the final conditions we can determine both θ_b and k solving this system of non-linear equation in such variables:

$$\begin{cases} x_b = \frac{1}{2}k^2(\theta_b - \sin\theta_b) \\ y_b = \frac{1}{2}k^2(1 - \cos\theta_b) \end{cases}$$

Chapter 4

Optimal Control Problem

The **optimal control problems** can be seen as a functional minimization, called **target**, subjected to ordinary differential constraints.

Isoperimetric problem Let's consider a function $y(x) \in [a,b] \to \mathbb{R}$ where the are known the point $y(a) = y(b) = 0$. If we consider the function as a rope having fixed length l that, by calculus I, can be computed as

$$l = \int_a^b \sqrt{1 - y'(x)^2}\, dx$$

an example of optimal control problem is the one to determine the function f that minimize it's integral over the it's domain:

$$\begin{aligned} \text{minimize:} \quad & \mathcal{F}(y) = -\int_a^b y(x)\, dx \\ \text{subject to:} \quad & y(a) = 0, y(b) = 0 \\ & \int_a^b \sqrt{1 + y'(x)^2}\, dx - l = 0 \end{aligned}$$

Discretized solution A way to solve the problem is by discretization, dividing the domain $[a,b]$ in n edges having length $h = \frac{b-a}{n}$ and so $x_k = a + kh$. By founding the discrete solution y_k of the system we can approximate also the continuous solution as

$$y_k \approx y(x_k)$$

Considering the integral approximation

$$\int_{x_{k-1}}^{x_k} y(x)\, dx \approx h y_{k-\frac{1}{2}} = h\frac{y_k + y_{k-1}}{2}$$

and the derivative

$$y'(x_k) \approx y'_{k-\frac{1}{2}} = \frac{y_k - y_{k-1}}{2}$$

With this being stated the initial isoperimetric problem can be rewritten (with some analytical simplification for the calculus) in the discretized form as

$$\begin{aligned} \text{minimize:} \quad & -\frac{h}{2}(y_0 + y_n) - h\sum_{k=1}^{n-1} y_k \\ \text{subject to:} \quad & y_0 = 0, y_n = 0 \\ & l - h\sum_{k=1}^{n} \sqrt{1 + \left(\frac{y_k - y_{k-1}}{h}\right)^2} \end{aligned}$$

As here stated, this problems become a constrained minimization one that can so be solved using the Lagrange multiplier method on which the unknowns is the vector of the discretized function $\boldsymbol{y} = (y_1, \dots, y_n)$ and the 3 Lagrange multipliers $\lambda_0, \lambda_n, \lambda$ associated to the constraints:

$$\mathcal{L}(\boldsymbol{y}, \lambda_0, \lambda_n, \lambda) = -\frac{h}{2}\sum_{k=1}^{n}\left(y_k + y_{k-1}\right) - \lambda_0 y_0 - \lambda_n y_n - \lambda\left(l - h\sum_{k=1}^{n}\sqrt{1 + \left(\frac{y_k - y_{k-1}}{h}\right)^2}\right)$$

By calculating the gradient of the lagrangian and setting it to zero, we satisfy the first necessary condition for the solutions that relates to the following non-linear system that can be solved by approximate solutions:

$$\begin{cases} -\frac{h}{2} - \lambda_0 + \lambda\frac{\frac{y_1 - y_0}{h}}{\sqrt{1 - \left(\frac{y_1 - y_0}{h}\right)^2}} = 0 & k = 0 \\ -h + \frac{\lambda\frac{y_k - y_{k+1}}{h}}{\sqrt{1 + \left(\frac{y_k - y_{k-1}}{h}\right)^2}} - \frac{\lambda\frac{y_{k+1} - y_k}{h}}{\sqrt{1 + \left(\frac{y_{k+1} - y_k}{h}\right)^2}} = 0 & \text{for } k = 1, \dots, n-1 \\ -\frac{h}{2} - \lambda_n - \lambda\frac{\frac{y_n - y_0}{h}}{\sqrt{1 + \left(\frac{y_1 - y_0}{h}\right)^2}} = 0 & k = n \\ y_0 = 0, \qquad y_n = 0 & \\ l - h\sum_{k=1}^{n}\sqrt{1 + \left(\frac{y_k - y_{k-1}}{h}\right)^2} = 0 & \end{cases}$$

Continuous interpretation Considering now the limit for $h \to 0$ the discretized systems seems to become continuous; in particular the 4-th conditions become

$$y_0 \to y(a) = 0 \qquad \text{and} \qquad y_n \to y(b) = 0$$

The first condition (associated to $k = 0$) and the third ($k = n$) instead becomes

$$-\frac{h}{2} - \lambda_0 + \lambda\frac{\frac{y_1 - y_0}{h}}{\sqrt{1 - \left(\frac{y_1 - y_0}{h}\right)^2}} \to -\lambda_0 - \lambda\frac{y'(a)}{\sqrt{1 - y'(a)^2}} = 0$$

$$-\frac{h}{2} - \lambda_n - \lambda\frac{\frac{y_n - y_0}{h}}{\sqrt{1 + \left(\frac{y_1 - y_0}{h}\right)^2}} \to -\lambda(b) - \lambda\frac{y'(b)}{\sqrt{1 - y'(b)^2}} = 0$$

With some analytical manipulation the second lagrange condition can be stated as

$$-1 - \lambda\frac{d}{dx}\left.\frac{y'(x)}{\sqrt{1 + y'(x)^2}}\right|_{x = x_k} = 0$$

Observation By pushing the limit $h \to 0$ the discretized problem becomes continuous and can relate to the general formulation of minimization of a functional $\mathcal{F}$ subjected to an equality constraints (depending on $\mathcal{G}$):

$$\begin{aligned} &\text{minimize:} && \mathcal{F}(y) = -\int_a^b y(x)\,dx && \Rightarrow \int_a^b L(y, y', t)\,dt \\ &\text{subject to:} && y(a) = 0, y(b) = 0 && \\ & && \int_a^b \sqrt{1 + y'(x)^2}\,dx - l = 0 && \Rightarrow \int_a^b \mathcal{G}(y, y', y)\,dt = 0 \end{aligned}$$

In the isoperimetric problem we can determine the functions $L(y, y', x) = -y$ whose derivative are $\frac{\partial L}{\partial y} = -1$ and $\frac{\partial L}{\partial y'} = 0$. Using so the Euler-Lagrange approach we can see that it fails because

$$\frac{\partial L}{\partial y} - \frac{d}{dx}\frac{\partial L}{\partial y'} = -1 \neq 0$$

Considering instead now the definition of the constraint $\mathcal{G}(y, y', x) = \frac{l}{b-a} - \sqrt{1 + y'^2}$ we can build another lagrangian $\tilde{L}$ that also considers (with a Lagrange multiplier λ) the constraint:

$$\tilde{L}(y, y', x, \lambda) = L(y, y', x) - \lambda \mathcal{G}(y, y', x)$$

Applying the Euler-Lagrange method on this expression it's possible to get the same result achieved with discretization, in fact

$$\frac{\partial \tilde{L}}{\partial y} - \frac{d}{dx}\frac{\partial \tilde{L}}{\partial y'} = -1 - \lambda \frac{y'}{\sqrt{1 + y'^2}}$$

4.1 General formulation from Euler-Lagrange

Considering the optimal control problem as

$$\begin{aligned} &\text{minimize:} && \int_a^b L(x, x', t)\, dt \\ &\text{subject to:} && x(a) = x_a, x(b) = x_b \\ & && \int_a^b \mathcal{G}(x, x', t)\, dt = 0 \end{aligned} \tag{4.1}$$

a way to reach a solution is by constructing the lagrangian $\mathcal{L}$ defined as

$$\begin{aligned} \tilde{L}(x, x', t, \lambda) &= L(x, x', t) - \lambda \mathcal{G}(x, x', t) \\ \mathcal{L}(x, \lambda, \lambda_a, \lambda_b) &= \int_a^b \tilde{L}(x, x', t, \lambda)\, dt - \lambda_a x(a) - \lambda_b x(b) \end{aligned} \tag{4.2}$$

At this point the solution of the problem is the stationary *point* (that in reality is a function) x of the lagrangian, and this means setting to zero the Gateaux derivative of $\mathcal{L}$:

$$\begin{aligned} \delta\mathcal{L} = \int_a^b \left(\frac{\partial \tilde{\mathcal{L}}}{\partial x}\delta_x + \frac{\partial \tilde{\mathcal{L}}}{\partial x'}\delta'_x + \frac{\partial \tilde{\mathcal{L}}}{\partial \lambda}\delta_\lambda \right) dt \\ - \delta_{\lambda_a}\big(x(a) - x_a\big) - \lambda_a \delta_{x(a)} - \delta_{\lambda_b}\big(x(b) - x_b\big) - \lambda_b \delta_{x(b)} \end{aligned}$$

Performing the integration by part in order to remove the term δ'_x the derivative so becomes

$$\delta\mathcal{L} = \int_a^b A\, \delta_x\, dx - \delta_{\lambda_a}\big(x(a) - x_a\big) - \delta_{\lambda_b}\big(x(b) - x_b\big) + B\delta_{x(a)} + C\delta_{x(b)}$$

where

$$A = \frac{\partial \tilde{\mathcal{L}}}{\partial x} - \frac{d}{dt}\frac{\partial \tilde{\mathcal{L}}}{\partial x'} \qquad B = -\lambda_a - \left.\frac{\partial \tilde{\mathcal{L}}}{\partial y'}\right|_a \qquad C = -\lambda_b + \left.\frac{\partial \tilde{\mathcal{L}}}{\partial y'}\right|_b$$

At this point the overall solution of the minimization problem becomes the following ordinary differential equation system:

$$\begin{cases} \left(\frac{\partial L}{\partial x} - \lambda \frac{\partial \mathcal{G}}{\partial x}\right) - \frac{d}{dx}\left(\frac{\partial L}{\partial x'} - \lambda \frac{\partial \mathcal{G}}{\partial x'}\right) = 0 \\ x(a) = x_a, \quad x(b) = x_b \\ -\lambda_a - \left(\left.\frac{\partial L}{\partial x'}\right|_a - \lambda \left.\frac{\partial \mathcal{G}}{\partial x'}\right|_a\right) = 0 \\ -\lambda_b + \left(\left.\frac{\partial L}{\partial x'}\right|_b - \lambda \left.\frac{\partial \mathcal{G}}{\partial x'}\right|_b\right) = 0 \end{cases} \tag{4.3}$$

Ordinary differential equation constraint Let's consider the general formulation of the isoperimetric problem as

$$\begin{aligned} \text{minimize:} \quad & \mathcal{F}(x) = \int_a^b L(x, x', t)\, dt \\ \text{subject to:} \quad & w\big(x(a), x(b)\big) = 0 \\ & \int_a^b \mathcal{G}(x, x', t)\, dt = l \end{aligned}$$

the problem can be solved introducing the function $z(t)$ such that $z' = \mathcal{G}(x, x', t)$: this means that $z(a) = 0$ and $z(b) = l$ and so the problem if transformed, considering also the introduction of the relation $y = x'$, to the form

$$\begin{aligned} \text{minimize:} \quad & \mathcal{F}(x, y) = \int_a^b L(x, y, t)\, dt \\ \text{subject to:} \quad & w\big(x(a), x(b)\big) = 0 \\ & z(a) = 0, \quad z(b) = l \\ & x' = y \\ & z' = \mathcal{G}(x, y, t) \end{aligned}$$

With the problem as here stated we refer to x and z as **states**, while y is the **control**. We can so see that an optimal control problem can be easily transform into a boundary value problem that can be solved using the technique shown on functional minimization.

4.2 General formulation

Considering the problem

$$\begin{aligned} \text{minimize:} \quad & \underbrace{\overbrace{\phi\big(\boldsymbol{x}(a), \boldsymbol{x}(b)\big)}^{\text{Mayer}} + \overbrace{\int_a^b L(\boldsymbol{x}, \boldsymbol{u}, t)\, dt}^{\text{Lagrange}}}_{\text{Bolza}} && : \text{target} \\ \text{subject to:} \quad & \boldsymbol{x}' = \boldsymbol{f}\big(\boldsymbol{x}, \boldsymbol{u}, t\big) && : \text{dynamical system} \\ & \boldsymbol{b}\big(\boldsymbol{x}(a), \boldsymbol{x}(b)\big) = 0 && : \text{boundary conditions} \\ & \int_a^b \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{u}, t)\, dt = \boldsymbol{g}_0 && : \text{integral constraints} \end{aligned}$$

where $\boldsymbol{x}$ is the vector of states and $\boldsymbol{u}$ the vector of the controls. To solve this kind of problem the first thing to do is to remove the integral constraints transforming them in simple boundary conditions; in order to so is the one to consider $g(\boldsymbol{x}, \boldsymbol{u}, t) = z'$ (adding so a new equation in the relation associated to the dynamical system) as a new derivate variable, then by integration it's possible to see that

$$z(a) = 0 \qquad z(b) = g_0$$

With that said the minimization problem becomes

$$\begin{aligned} \text{minimize:} \quad & \phi\big(\boldsymbol{x}(a), \boldsymbol{x}(b)\big) + \int_a^b L(\boldsymbol{x}, \boldsymbol{u}, t)\, dt \\ \text{subject to:} \quad & \boldsymbol{x}' = \boldsymbol{f}\big(\boldsymbol{x}, \boldsymbol{u}, t\big) \\ & \boldsymbol{z}' = \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{u}, t) \\ & \boldsymbol{b}\big(\boldsymbol{x}(a), \boldsymbol{x}(b)\big) = 0 \\ & \boldsymbol{z}(a) = 0 \qquad \boldsymbol{z}(b) = \boldsymbol{g}_0 \end{aligned}$$

Compacting the vector of states $\boldsymbol{x}$ and added variable $\boldsymbol{z}$ we can define a new variable vector $\boldsymbol{w}$, and similarly all the dynamical system can be described by a multi-variable function F and boundary conditions B and so the problem can be compacted as

$$\begin{aligned} \text{minimize:} \quad & \psi\big(\boldsymbol{w}(a), \boldsymbol{w}(b)\big) + \int_a^b \mathcal{M}(\boldsymbol{w}, \boldsymbol{u}, t)\, dt \\ \text{subject to:} \quad & \boldsymbol{w}' = \boldsymbol{F}\big(\boldsymbol{w}, \boldsymbol{u}, t\big) \\ & \boldsymbol{B}\big(\boldsymbol{w}(a), \boldsymbol{w}(b)\big) = 0 \end{aligned}$$

To solve the problem as here state it's necessary to build the function of the boundary conditions $\boldsymbol{\mathcal{B}}$ (depending by the Lagrange multipliers $\boldsymbol{\mu}$), also known as **utility function**, and the **Hamiltonian** $\mathcal{H}$ (depending by the multipliers $\boldsymbol{\lambda}$):

$$\begin{aligned} \boldsymbol{\mathcal{B}}\big(\boldsymbol{w}(a), \boldsymbol{w}(b), \boldsymbol{\mu}\big) &= \psi\big(\boldsymbol{w}(a), \boldsymbol{w}(b)\big) + \boldsymbol{\mu} \cdot \boldsymbol{B}\big(\boldsymbol{w}(a), \boldsymbol{w}(b)\big) \\ \mathcal{H}(\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{u}, t) &= \mathcal{M}(\boldsymbol{w}, \boldsymbol{u}, t) + \boldsymbol{\lambda} \cdot \boldsymbol{F}(\boldsymbol{w}, \boldsymbol{u}, t) \end{aligned} \tag{4.4}$$

At this point it's possible to compute the lagrangian $\mathcal{L}$ on which the variations can be performed:

$$\begin{aligned} \mathcal{L}\big(\boldsymbol{w}, \boldsymbol{u}, \boldsymbol{\lambda}, \boldsymbol{\mu}, t\big) &= \boldsymbol{\mathcal{B}}\big(\boldsymbol{w}(a), \boldsymbol{w}(b), \boldsymbol{\mu}\big) + \mathcal{H}(\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{u}, t) \\ \delta\mathcal{L} &= \int_a^b \Big(A\delta_w + B\delta_u + C\delta_\lambda\Big)\, dt + D\,\delta_{w(a)} + E\,\delta_{w(b)} - \boldsymbol{B}\big(w(a), w(b)\big)\delta_\mu \end{aligned} \tag{4.5}$$

where

$$A = \lambda' + \frac{\partial \mathcal{H}}{\partial w} \qquad B = \frac{\partial \mathcal{H}}{\partial u} \qquad C = f(w, u, t) - x'$$

$$D = \left.\frac{\partial \mathcal{B}}{\partial w(a)}\right|_a + \lambda(a) \qquad E = \left.\frac{\partial \mathcal{B}}{\partial w(b)}\right|_b - \lambda(b)$$

With that said the resulting boundary valued problem becomes

$$\begin{cases} w' = f(x, u, t) & \text{: original ODE} \\ \lambda' = -\dfrac{\partial \mathcal{H}}{\partial w} & \text{: adjoint ODE} \\ B\big(w(a), w(b)\big) = 0 & \text{: original boundary condition} \\ \left.\dfrac{\partial \mathcal{B}}{\partial w(a)}\right|_a + \lambda(a) \\ \left.\dfrac{\partial \mathcal{B}}{\partial w(a)}\right|_b - \lambda(b) \end{cases} \text{: adjoint boundary conditions} \\ \dfrac{\partial \mathcal{H}}{\partial u} = 0 & \text{: control equation} \end{cases} \tag{4.6}$$

Example 4.1: optimal control problem

Let's consider the problem of a mass m sliding on a plane (coordinate x) subjected only to an external applied force F, the optimal control problem is

$$\begin{aligned} \text{minimize:} \quad & \int_0^1 F^2\, dt \\ \text{subject to:} \quad & x' = v, \qquad v' = \frac{F}{m} \\ & x(0) = 0 \qquad x(1) = 1 \\ & v(0) = 0 \qquad v(1) = 0 \end{aligned}$$

In this case the description of the dynamical system is known from the physical domain, in fact

the velocity is the derivative of the position and the acceleration (derivative of velocity) is equal to the force applied divided by the mass; the boundaries conditions are set by the problem.
At this point to solve the problem we have to compute the two function $\mathcal{B}, \mathcal{H}$ defined as

$$\mathcal{B} = \mu_1\big(x(a) - 0\big) + \mu_2\big(x(b) - 1\big) + \mu_3\big(v(a) - 0\big) + \mu_4\big(v(b) - 0\big)$$
$$\mathcal{H} = F^2 + \lambda_1 v + \lambda_2 \frac{F}{m}$$

Following the results of equation 4.6 the boundary valued problem that has to be solved to find the minimum point is

$$\begin{cases} x' = v \\ v' = F/m \\ \lambda_1' = 0 \\ \lambda_2' = \lambda_1 \\ x(0) = 0 \qquad x(1) = 1 \\ v(0) = 0 \qquad v(1) = 0 \\ \mu_1 + \lambda_1(0) = 0 \\ \mu_3 + \lambda_2(0) = 0 \\ \mu_2 - \lambda_1(1) = 0 \\ \mu_4 - \lambda_2(1) = 0 \\ 2F + \frac{\lambda_2}{m} = 0 \end{cases}$$

In this case the adjoin boundary conditions are trivially solved (in fact will exists μ_i that will satisfy the equation); considering the last equation in the expression of the velocity the system becomes

$$\begin{cases} x' = v \\ v' = -\frac{\lambda_2}{2m^2} \\ \lambda_1' = 0 \\ \lambda_2' = \lambda_1 \\ x(0) = 0 \qquad x(1) = 1 \\ v(0) = 0 \qquad v(1) = 0 \end{cases}$$

Considering that λ_1 is a constant c_1 (in order to have null derivative), then it means that λ_2 is in the form $c_1 t + c_2$ and so the expression of the acceleration and velocity becomes

$$v' = -\frac{c_1 t + c_2}{2m^2} \qquad \xrightarrow{\int} \qquad v = -\frac{c_1}{4m^2} t^2 - \frac{c_2}{2m^2} t + c_3$$

$$\xrightarrow{\int} \qquad x = -\frac{c_1}{12m^2} t^3 - \frac{c_2}{4m^2} t^2 + c_3 t + c_4$$

Considering the boundary conditions than the integration constant are $c_3 = c_4 = 0$ and the solution of the linear system given by $-c_1 - 2c_2 = 0$ and $-c_1 - 3c_2 = 12m^2$ that determines the last full solution

$$x = -6t^2 + 6t \qquad v = -2t^3 + 3t^2 \qquad \lambda_1 = 24m^2 \qquad \lambda_2 = 24m^2 t - 12m^2$$

and so the control history to determine this minimal solution is

$$F = -\frac{\lambda_2}{2m} = 6\big(1 - 2t\big)m$$

4.3 Free time problem

Let's consider the problem on which the time T is unknown and so is a variable:

$$\begin{aligned} \text{minimize:} \quad & \phi\big(x(0), x(T)\big) + \int_0^T L(x,u,t)\,dt \\ \text{subject to:} \quad & x' = f(x,u,t) \\ & b\big(x(0), x(T)\big) = 0 \end{aligned}$$

To solve this kind of problem it's possible to perform a change of variable $t = sT$ in order to have minimization interval bounded from 0 to 1; with that said we can define the state $\tilde{x}(s) = x(sT)$ (and so $\tilde{x}'(s) = x'(sT)T$) and the control $\tilde{u}(s) = u(sT)$; the dynamical system $x'(sT) = f\big(x(sT), u(sT), sT\big)$ can be rewritten as $\tilde{x}'(s) = \tilde{f}(\tilde{x}, \tilde{u}, T, s)$ where $\tilde{f}(x,u,T,s) = T\,f(x,u,sT)$. Similarly the lagrangian running equation becomes $\tilde{L}(x,u,T,s) = TL(x,u,sT)$.

In general the free time problem, with the changed variable, can be solved as

$$\begin{aligned} \text{minimize:} \quad & \phi\big(\tilde{x}(0), \tilde{x}(T)\big) + \int_0^1 TL(\tilde{x}, \tilde{u}, T\,s)\,ds \\ \text{subject to:} \quad & \tilde{x}' = T\,f(\tilde{x}, \tilde{u}, Ts) \\ & b\big(\tilde{x}(0), \tilde{x}(1)\big) = 0 \end{aligned}$$

4.4 Pontryagin maximum (minimum) principle

The **Pontryagin maximum** (minimum) **principle** has been developed in order to solve problem where the controls must stay bounded to a certain range due to the physical implementation of the system (if, as example, a motor is powered with $20W$ of power, than it cannot move object that require more than that power).

Mathematically this kind of problem is described by

$$\begin{aligned} \text{minimize:} \quad & \phi\big(x(a), x(B)\big) + \int_a^b L(x,u,t)\,dt \\ \text{subject to:} \quad & x' = f(x,u,t) \\ & b\big(x(a), x(b)\big) = 0 \\ & u(t) \in \mathcal{U} \end{aligned}$$

where $\mathcal{U}$ is the **domain of the controls** that, in order for the principle to work, must be convex and compact.

To solve this problem we compute the Hamiltonian $\mathcal{H}$ and the utility function $\mathcal{B}$ as in the previous cases:

$$\mathcal{H}(x,u,\lambda,t) = L(x,u,t) + \lambda f(x,u,t) \qquad \mathcal{B}(x_a, x_b, \mu) = \phi(x_a, x_b) + \mu b(x_a, x_b)$$

With that stated the boundary valued problem becomes similar to the formulation in equation 4.6

(page 61) with the addition of the Pontryagin minimum principle:

$$\begin{cases} x' = f(x,u,t) = \dfrac{\partial \mathcal{H}}{\partial \lambda} & : \text{original ODE} \\ \lambda' = -\dfrac{\partial \mathcal{H}}{\partial x} & : \text{adjoint ODE - co-equations} \\ b\big(w(a), w(b)\big) = 0 & : \text{original BC} \\ \left.\begin{array}{l} \dfrac{\partial \mathcal{B}}{\partial x_a} + \lambda(a) \\ \dfrac{\partial \mathcal{B}}{\partial x_b} - \lambda(b) \end{array}\right\} & : \text{adjoint BC} \\ u(t) = \underset{\overline{u}\in\mathcal{U}}{\operatorname{argmin}}\big\{\mathcal{H}(x,\overline{u},\lambda,t)\big\} & : \text{Pontryagin min principle} \\ \dfrac{\partial \mathcal{H}}{\partial u} = 0 & : \text{control equation} \end{cases} \tag{4.7}$$

Example 4.2: maximum travel optimal control problem

Let's consider the system in example 4.1 (page 61) where in this case the goal is to maximize the travel of the mass m having a force $F \leq |1|$; this means solving the following optimal control problem

$$\begin{array}{ll} \text{minimize:} & -x(1) \\ \text{subject to:} & x' = v, \qquad v' = \dfrac{F}{m} \\ & x(0) = 0 \\ & v(0) = 0 \qquad v(1) = 0 \\ & F \in [-1,1] \end{array}$$

In order to solve this problem it's necessary to determine the Hamiltonian $\mathcal{H}(x,v,\lambda_1,\lambda_2,F,t) = \lambda_1 v + \lambda_2 \frac{F}{m}$ and the utility function $\mathcal{B}(x_a,v_a,x_b,v_b,\mu_1,\mu_2,\mu_3) = -x_b + \mu_1 x_a + \mu_2 v_a + \mu_3 v_b$. The adjoint equation so becomes $\lambda_1' = -\frac{\partial \mathcal{H}}{\partial x} = 0$ and $\lambda_2' = -\frac{\partial \mathcal{H}}{\partial v} = -\lambda_1$; for the the adjoint boundary conditions only $-1 - \lambda_1(1) = 0$ is reported (the other equation trivilly solves the cofficients μ_i that gives no information for the final solution). The boundary valued problem so become

$$\begin{cases} x' = v \\ v' = F/m \\ \lambda_1' = 0 \\ \lambda_2' = -\lambda_1 \\ -1 - \lambda_1(1) = 0 \\ x(0) = v(0) = v(1) = 0 \\ F(t) = \underset{\overline{F}\in[-1,1]}{\operatorname{argmin}} \left\{ \lambda_1(t) v + \lambda_2(t) \dfrac{\overline{F}}{m} \right\} \end{cases}$$

In this case in order to compute the argument minimum associated to the Pontryagin principle

we have to consider only the term related to $\overline{F}$, and in particular this means

$$\begin{aligned} F(t) &= \underset{\overline{F}\in[-1,1]}{\operatorname{argmin}} \left\{ \lambda_2(t)\frac{\overline{F}}{m} \right\} \\ &= \begin{cases} +1 & \text{if } \lambda_2(t) < 0 \\ -1 & \text{if } \lambda_2(t) > 0 \\ [-1,1] & \text{if } \lambda_2(t) = 0 \end{cases} \\ &= -\operatorname{sign}\lambda_2(t) \end{aligned}$$

and so the simplified system that determines the solution becomes

$$\begin{cases} x' = v \\ v' = -\dfrac{\operatorname{sign}\lambda_2}{m} \\ \lambda_1' = 0 \\ \lambda_2' = -\lambda_1 \\ \lambda_1(1) = -1 \\ x(0) = v(0) = v(1) = 0 \end{cases}$$

By integration it can be seen that λ_1 can be regarded as a constant c_1 that, for the adjoint boundary condition, is equal to -1 and so λ_2 is in the form

$$\lambda_2 = t + c_2$$

From now on the solution becomes analytically complex, however it can be found that $c_2 = -\frac{1}{2}$ determining the solution

$$F(t) = \begin{cases} 1 & t \leq \frac{1}{2} \\ -1 & t > \frac{1}{2} \end{cases}$$

Example 4.3: optimal control problem

Given the optimal control problem

$$\begin{aligned} &\text{minimize:} && \mathcal{F}(x,u) = \int_0^2 2x - u^2\, dt \\ &\text{subject to:} && x' = 1 - 2u \\ & && x(0) = 1 \qquad\qquad x(2) = 0 \end{aligned}$$

the solution can be obtained by computing the hamiltonian $\mathcal{H}$ and the utility function $\mathcal{B}$ related to the problem:

$$\begin{aligned} \mathcal{H}(x,u,\lambda) &= 2x - u^2 - \lambda(1-2u) \\ \mathcal{B}(x_0,x_2,\mu_1,\mu_2) &= -\mu_1(x_0 - 1) - \mu_2 x_2 \end{aligned}$$

With that stated, the lonely resulting co-equation is $\lambda' = \frac{\partial \mathcal{H}}{\partial x} = 2$, determining so that the solution of $\lambda(t)$ is in the form $2t + c_1$ (where c_1 is a constant depending from the boundary conditions). The adjoint boundary conditions aren't useful to determine the solution because they trivially solves the parameters μ_1, μ_2. We still have the control equation that states

$$\frac{\partial \mathcal{H}}{\partial u} = -2u + 2\lambda = 0 \qquad \Rightarrow \quad u(t) = \lambda(t)$$

Using the original ordinary differential equation we have that $x' = 1 - 4t - 2c_1$ that by integration becomes

$$x(t) = -2t^2 + (1 - 2c_1)t + c_2$$

From the boundary condition $x(0) = 1$ we obtain $c_2 = 1$, while considering $x(2) = 0$ we have $c_1 = -5/4$ and so the final control law is

$$u(t) = 2t - \frac{5}{4}$$

Example 4.4: optimal control problem

To solve the optimal control problem

$$\begin{aligned}
&\text{minimize:} && \int_0^1 t + x - y^2\, dt \\
&\text{subject to:} && x' = x + y \qquad y' = x + yu \\
& && \int_0^1 \sin y\, dt = 3 \\
& && x(0) = 1 \qquad y(1) = 0 \\
& && u \in [-2, 3]
\end{aligned}$$

the first thing to do is removing the integral constraint by transforming him in a differential equation with 2 boundary condition as follows:

$$\begin{aligned}
&\text{minimize:} && \int_0^1 t + x - y^2\, dt \\
&\text{subject to:} && x' = x + y \qquad y' = x + yu \qquad z' = \sin y \\
& && x(0) = 1 \qquad y(1) = 0 \qquad z(0) = 0 \qquad z(1) = 3 \\
& && u \in [-2, 3]
\end{aligned}$$

The associated hamiltonian $\mathcal{H}$ and utility function $\mathcal{B}$ of the problem are

$$\begin{aligned}
\mathcal{H}(x, y, z, \lambda_1, \lambda_2, \lambda_3) &= t + x - y^2 - \lambda_1(x + y) - \lambda_2(x + yu) - \lambda_3(\sin y) \\
\mathcal{B}(x_0, y_1, z_0, z_1, \mu_{1\ldots4}) &= -\mu_1(x_0 - 1) - \mu_2 y_1 - \mu_3 z_0 - \mu_4(z_1 - 3)
\end{aligned}$$

With that stated the co-equations are computed as

$$\begin{aligned}
\lambda_1' &= -\frac{\partial \mathcal{H}}{\partial x} = -1 + \lambda_1 + \lambda_2 \\
\lambda_2' &= -\frac{\partial \mathcal{H}}{\partial y} = 2y + \lambda_1 + \lambda_2 u + \lambda_3 \cos y \\
\lambda_3' &= -\frac{\partial \mathcal{H}}{\partial z} = 0
\end{aligned}$$

Evaluating the adjoint boundary conditions for the point that present no explicit constraints (so x_1, y_0) we retrieve $\lambda_1(1) = 0$ and $\lambda_2(0) = 0$. With the definition of the control equation $\frac{\partial \mathcal{H}}{\partial u} = -\lambda_2 y$

the resulting boundary value problem of the optimal control one is

$$\begin{cases} x' = x + y \\ y' = x + yu \\ z' = \sin y \\ \lambda_1' = \lambda_1 + \lambda_2 - 1 \\ \lambda_2' = 2y + \lambda_1 + \lambda_2 y + \lambda_3 \cos y \\ \lambda_3' = 0 \\ x(0) = 1 \qquad \lambda_1(1) = 0 \\ y(1) = 0 \qquad \lambda_2(0) = 0 \\ z(0) = 0 \qquad z(1) = 3 \\ -\lambda_2 = 0 \\ u = \underset{\overline{u} \in [-2,3]}{\operatorname{argmin}} \{-\lambda_2 y \overline{u}\} \end{cases}$$

where the solution of the Pontryagin minimum principle is

$$u(t) = \begin{cases} -2 & \lambda_2 y < 0 \\ 3 & \lambda_2 y \geq 0 \end{cases}$$

Example 4.5: optimal control problem

Given the optimal control problem

$$\begin{aligned} \text{minimize:} \quad & \mathcal{F}(x,y,u) = x(0) + \int_0^1 x^2 + y^2 + u \, dt \\ \text{subject to:} \quad & x' = y - u \qquad y' = x + u \\ & x(1) = 2 \\ & u(t) \in [-1,2] \end{aligned}$$

the solution can be obtained by firstly computing the hamiltonian $\mathcal{H}$ and the utility function $\mathcal{B}$ of the problems resulting in

$$\begin{aligned} \mathcal{H}(x,y,u,\lambda_1,\lambda_2) &= x^2 + y^2 + u + \lambda_1(y-u) + \lambda_2(x+u) \\ &= x^2 + y^2 + u(1 - \lambda_1 + \lambda_2) + \lambda_1 y + \lambda_2 x \\ \mathcal{B}(x_0, x_1, \mu) &= x_0 + \mu(x_1 - 2) \end{aligned}$$

We can so explicit: the adjoint ordinary differential equations $\lambda_1' = -\frac{\partial \mathcal{H}}{\partial x} = -2x - \lambda_2$, $\lambda_2' = -\frac{\partial \mathcal{H}}{\partial y} = -2y - \lambda_1$, the adjoint boundary conditions $\lambda_1(0) = -\frac{\partial \mathcal{B}}{\partial x_0} = -1$, $\lambda_1(1) = \frac{\partial \mathcal{B}}{\partial x_1} = \mu$, $\lambda_2(0) = -\frac{\partial \mathcal{B}}{\partial y_0} = 0$, $\lambda_2(1) = \frac{\partial \mathcal{B}}{\partial y_1} = 0$ and the control equation $\frac{\partial \mathcal{H}}{\partial u} = 1 - \lambda_1 + \lambda_2 = 0$. With the

addition of the original differential equation the boundary value problem becomes

$$\begin{cases} x' = y - u \\ y' = x + u \\ \lambda_1' = -2x - \lambda_2 \\ \lambda_2' = -2y - \lambda_1 \\ \lambda_1(0) = -1 \\ \cancel{\lambda_1(1) = \mu} & : \text{trivially satisfied} \\ \lambda_2(0) = \lambda_2(1) = 0 \\ 1 + \lambda_1 - \lambda_2 = 0 \\ u = \underset{\overline{u} \in [-1,2]}{\operatorname{argmin}} \left\{ \overline{u}\left(1 + \lambda_1 - \lambda_2\right) \right\} \end{cases}$$

In this case the solution of the Pontryagin minimum principle is reduced to the form

$$u(t) = \begin{cases} -1 & 1 + \lambda_1 - \lambda_2 > 0 \\ 2 & 1 + \lambda_1 - \lambda_2 < 0 \end{cases}$$

Example 4.6: free time problem

Given the free time optimal control problem

$$\begin{aligned} \text{minimize:} \quad & \int_0^T \left(1 + t^2\right) u^2 + y \, dt \\ \text{subject to:} \quad & x' = y + u \qquad y' = xy \\ & x(0) = 0 \qquad y(T) = 1 \end{aligned}$$

the first thing to do is to rewrite the statement in a form where the integral bound are known by performing a change of variable ζ that we arbitrary set to be defined in the domain $[0, 1]$ resulting in the transformation $\zeta = t/T$. Following the results explained at page 63 the optimal control problem can so be stated as

$$\begin{aligned} \text{minimize:} \quad & \int_0^1 T\Big(\left(1 + \zeta^2 T^2\right) u^2(\zeta) + y(\zeta) \Big) \, d\zeta \\ \text{subject to:} \quad & x'(\zeta) = T\Big(y(\zeta) + u(\zeta)\Big) \qquad y' = Tx(\zeta)y(\zeta) \\ & T'(\zeta) = 0 \\ & x(0) = 0 \qquad y(1) = 1 \end{aligned}$$

The hamiltonian $\mathcal{H}$ related to such problem is so

$$\mathcal{H} = T\left(1 + \zeta^2 T^2\right) u^2 + Ty - \lambda_1 T\left(y + u\right) - \lambda_2 Txy + 0\lambda_3$$

Considering the original ordinary differential equation of the problem and the addition of the co-equations the system that needs to be solved is

$$\begin{cases} x'(\zeta) = T\Big(y(\zeta) + u(\zeta)\Big) \\ y' = Tx(\zeta)y(\zeta) \\ T'(\zeta) = 0 \\ \lambda_1'(\zeta) = -\frac{\partial \mathcal{H}}{\partial x} = \lambda_2 Ty(\zeta) \\ \lambda_2'(\zeta) = -\frac{\partial \mathcal{H}}{\partial y} = -T + \lambda_1 + \lambda_2 Tx(\zeta) \\ \lambda_3'(\zeta) = -\frac{\partial \mathcal{H}}{\partial T} = -u^2(\zeta) - 3\zeta^2 T^2 u^2(\zeta) - y(\zeta) - \lambda_1\Big(y(\zeta) + u(\zeta)\Big) - \lambda_2 x(\zeta)y(\zeta) \end{cases}$$

on which the boundary conditions are

$$x(0) = 0 \qquad \lambda_1(1) = 0 \qquad \lambda_2(0) = 0 \qquad y(1) = 1 \qquad \lambda_3(0) = \lambda_3(1) = 0$$

We also need to consider the control law $\frac{\partial \mathcal{H}}{\partial u}$ that becomes $2T\big(1+\zeta^2 T^2\big)u - \lambda_1 = 0$ and so we can explicit the control u as

$$u(\zeta) = \frac{\lambda_1(\zeta)}{2T\big(1+\zeta^2 T^2\big)}$$

Example 4.7: exam's simulation

Given the optimal control problem in the control u and states x, y

$$\begin{aligned} \text{minimize:} \quad & \int_0^1 \big(1+x^2\big)u^2\,dt \\ \text{subject to:} \quad & x' = xy - u \qquad y' = x + u \\ & \int_0^1 \big(x+u\big)\,dt = 2 \\ & y(0) = 2 \end{aligned}$$

the first thing to do is remove the integral constraint by adding a new state variable z such that $z' = x + u$ with boundary conditions $z(0) = 0$ and $z(1) = 2$; with that the problem becomes

$$\begin{aligned} \text{minimize:} \quad & \int_0^1 \big(1+x^2\big)u^2\,dt \\ \text{subject to:} \quad & x' = xy - u \qquad y' = x + u \qquad z' = x + u \\ & y(0) = 2 \qquad z(0) = 0 \qquad z(1) = 2 \end{aligned}$$

We can so compute the hamiltonian $\mathcal{H}$ and the utility function $\mathcal{B}$ of the problem as

$$\begin{aligned} \mathcal{H}(x, y, z, u, \lambda_1, \lambda_2, \lambda_3, t) &= \big(1+x^2\big)u^2 + \lambda_1\big(xy - u\big) + \lambda_2\big(x+u\big) + \lambda_3\big(x+u\big) \\ \mathcal{B}(y_0, z_0, z_1, \mu_1, \mu_2, \mu_3) &= \mu_1\big(y_0 - 2\big) + \mu_2 z_0 + \mu_3\big(z_1 - 2\big) \end{aligned}$$

The control obtained by the Pontryagin maximum principle is the solution of u that verifies $\frac{\partial \mathcal{H}}{\partial u} = 0$, and so we have

$$\frac{\partial \mathcal{H}}{\partial u} = 2\big(1+x^2\big)u - \lambda_1 + \lambda_2 + \lambda_3 \qquad \Rightarrow \quad u(t) = \frac{\lambda_1(t) - \lambda_2(t) - \lambda_3(t)}{2\big(1+x^2\big)}$$

The additional adjoint ODE are described by the following derivatives:

$$\begin{aligned} \lambda_1' &= -\frac{\partial \mathcal{H}}{\partial x} = -2u^2 x - \lambda_1 y - \lambda_2 \\ \lambda_2' &= -\frac{\partial \mathcal{H}}{\partial y} = -\lambda_1 x \\ \lambda_3' &= -\frac{\partial \mathcal{H}}{\partial z} = 0 \end{aligned}$$

while the adjoint boundary conditions are

$$\lambda_1(0) = -\frac{\partial \mathcal{B}}{\partial x_0} = 0 \qquad \lambda_1(1) = \frac{\partial \mathcal{B}}{\partial x_1} = 0$$

$$\cancel{\lambda_2(0) = -\frac{\partial \mathcal{B}}{\partial y_0} = -\mu_1} \qquad \lambda_2(1) = \frac{\partial \mathcal{B}}{\partial y_1} = 0$$

$$\cancel{\lambda_3(0) = -\frac{\partial \mathcal{B}}{\partial z_0} = -\mu_2} \qquad \cancel{\lambda_3(1) = \frac{\partial \mathcal{B}}{\partial z_1} = \mu_3}$$

Some expression have been cancelled out because they are trivially solved for μ_i.

Example 4.8: free time problem, 2 controls

Given the free time optimal control problem in the controls u, w and states x, y

$$\begin{aligned}
\text{minimize:} \quad & \int_0^T \Big(x^2(t) + y^2(t)\Big)\, dt \\
\text{subject to:} \quad & x(t)' = y(t)u(t) \qquad y'(t) = x(t) + w(t) \\
& x(0) = 0 \qquad y(0) = 0 \qquad y(T) = 1 \\
& u(t), w(t) \in [-1, 1]
\end{aligned}$$

the first thing to find the solution is to transform the problem in fixed boundary optimal control problem by performing the change of variable $t = \zeta T$ (with $z \in [0, 1]$) and so

$$\begin{aligned}
\text{minimize:} \quad & \int_0^1 T(\zeta)\Big(x^2(\zeta) + y^2(\zeta)\Big)\, d\zeta \\
\text{subject to:} \quad & x(\zeta)' = T(\zeta)y(\zeta)u(\zeta) \qquad y'(\zeta) = T(\zeta)\Big(x(\zeta) + w(\zeta)\Big) \qquad T'(\zeta) = 0 \\
& x(0) = 0 \qquad y(0) = 0 \qquad y(1) = 1 \\
& u(\zeta), w(\zeta) \in [-1, 1]
\end{aligned}$$

From now on all the functions will be represented as depending from the independent variable ζ and it won't be written any more. With that said we can compute the hamiltonian $\mathcal{H}$ and the utility function $\mathcal{B}$ of the problem as

$$\begin{aligned}
\mathcal{H}(x, y, u, w, \lambda_1, \lambda_2, \lambda_3, T) &= T\big(x^2 + y^2\big) + \lambda_1 Tyu + \lambda_2 T\big(x + w\big) + 0 \cdot \lambda_3 \\
\mathcal{B}(x_0, y_0, y_1, \mu_1, \mu_2, \mu_3) &= \mu_1 x_0 + \mu_2 y_0 + \mu_3\big(y_1 - 1\big)
\end{aligned}$$

We that stated we can build the boundary value problem solution of the free time optimal control problem by adding at the initial ODEs and boundary conditions the the adjoint ones and the control law (related to Pontryagin minimum principle):

$$\begin{cases}
x' = Tyu \\
y' = T(x + w) \\
T' = 0 \\
\lambda_1' = -\frac{\partial \mathcal{H}}{\partial x} = -2xT - \lambda_2 T \\
\lambda_2' = -\frac{\partial \mathcal{H}}{\partial y} = -2Ty - \lambda_1 Tu \\
\lambda_3' = -\frac{\partial \mathcal{H}}{\partial T} = -x^2 - y^2 - \lambda_1 yu - \lambda_2(x + w) \\
x(0) = 0 \qquad y(0) = 0 \qquad y(1) = 1 \\
\lambda_1(1) = \lambda_3(0) = \lambda_3(1) = 0 \\
\frac{\partial \mathcal{H}}{\partial u} = \lambda_1 Ty = 0 \\
\frac{\partial \mathcal{H}}{\partial w} = \lambda_2 T = 0 \\
u = \underset{\overline{u} \in [-1,1]}{\operatorname{argmin}} \{\lambda_1 Ty\overline{u}\} \\
w = \underset{\overline{w} \in [-1,1]}{\operatorname{argmin}} \{\lambda_2 T\overline{w}\}
\end{cases}$$

The solution of the control obtained by the Pontryagin principle becomes

$$u(\zeta) = \begin{cases} -1 & \lambda_1 Ty > 0 \\ 1 & \lambda_1 Ty < 0 \end{cases} \qquad w(\zeta) = \begin{cases} -1 & \lambda_2 T > 0 \\ 1 & \lambda_2 T < 0 \end{cases}$$

The solution of the boundary value problem are in the independent variable ζ, so to have the final solution in the time variable t every control/state law must be evaluated in $t = \zeta T$ (where T has been obtained by the solution of the boundary value problem).

Part II

Differential Algebraic Equations

Chapter 5

Ordinary Differential Equations and Numerical Solutions

To start the description of the **differential algebraic equations** **DAEs** it's firstly necessary to recall what **ordinary differential equations** **ODEs** are, what they mean and how to solve them.

In particular ordinary differential equations is a particular equation that involves a function (for example $y(x)$ depending from the independent variable x) and it's derivative as shown in this example:

$$y''(x) + xy'(x) + y(x) = \sin x \qquad \leftrightarrow \qquad y'' + xy' + y = \sin x$$

Ordinary differential equations (or system of ODEs) can be written in a ***standard form*** made by the **differential part**, where the first derivative is a function of itself, and the **initial condition** that set a specified value of the solution of the problem:

$$\begin{cases} y' = f(x, y(x)) = f(x, y) & : \text{differential part} \\ y(a) = y_a & : \text{initial condition} \end{cases} \tag{5.1}$$

Initial conditions are mandatory: the solution of the differential part lonely gives a *family* of solutions (parametric results) whose specific value can be determined only by knowing the value of the function at certain time. Considering the simple case of the differential $x' = 0$ with independent variable t, then the general solution is the class of all the constant functions $x(t) = c$ (with $c \in \mathbb{R}$). If a boundary condition is set (example $x(1) = 3$) then we can chose the particular solution (in this case $c = 3$ and so $x(t) = 3$).

Ordinary differential equations can also come in system as in the following example (with t as dependent variable) composed by 2 differential and 2 initial condition terms:

$$\begin{cases} x' = x + y \\ y' = e^x - y \\ x(0) = 0 \\ y(0) = 1 \end{cases}$$

Vectorial notation Considering the system of $n = 3$ differential equation depending from the independent variable t in the form

$$\begin{cases} z'(t) = x(t) + z(t) \\ w'(t) = z(t) \\ x'(t) = w(t)z(t) + t(t) \\ x(0) = w(0) = z(0) = 1 \end{cases}$$

then it can be rewritten in a vectorial form; considering in fact the substitutions $x(t) = y_1(t)$, $w(t) = y_2(t)$ and $z(t) = y_3(t)$ we have obtain the system

$$\begin{cases} y_3' = y_1 + y_3 \\ y_2' = y_3 \\ y_1' = y_2 y_3 + t \\ y_1(0) = y_2(0) = y_3(0) = 1 \end{cases} \Rightarrow \begin{cases} y_1' = y_2 y_3 + t \\ y_2' = y_3 \\ y_3' = y_1 + y_3 \\ y_1(0) = y_2(0) = y_3(0) = 1 \end{cases}$$

Considering the vector $\boldsymbol{y} = (y_1, y_2, y_3)$ we can so rewrite the original system of ODEs as

$$\begin{cases} \boldsymbol{y}' = \boldsymbol{f}(t, \boldsymbol{y}) \\ \boldsymbol{y}(0) = \mathbf{1} \end{cases} \tag{5.2}$$

where

$$\boldsymbol{f} : \mathbb{R}^{n+1} \to \mathbb{R}^n = \begin{pmatrix} y_2 y_3 + t \\ y_3 \\ y_1 + y_3 \end{pmatrix}$$

In this case the domain of the function $\boldsymbol{f}$ is a vector of dimension $n + 1$: n related to the number of ODEs and 1 for the time dependency.

Order of an ODE The **order** of an ordinary differential equation is equal to the maximum order derivative appearing in the differential part of the system; considering as example

$$\begin{cases} y'' + y' + z' = 0 \\ z' + t = 0 \end{cases}$$

the order of such ordinary differential system is equal to 2 (associated to the term y'').

In general numerical methods are defined for 1^{st} order ODEs and so it's necessary (but most importantly possible) to convert any generic differential equation into a system of 1^{st} order ODEs by performing a change of variable.

Example 5.1: reduction to a system of ODEs of first order

Given the system of ODEs of the 3^{rd} order in the independent variable t defined as

$$\begin{cases} x''' + y' = x^2 + t \\ y'' + x = t^2 + 1 \\ x(0) = 0 \qquad y(0) = 0 \\ x'(0) = 0 \qquad y'(0) = 2 \\ x''(0) = 2 \end{cases}$$

the reduction to a system of first order ODEs is made by introducing the variable $z = x'$; defining instead the function $x'' = z' = w$ we also have that $z' = w$ hence $x''' = z'' = w'$; finally we can set

$y' = p$ hence $y'' = p$. The system so becomes

$$\begin{cases} w' + p = x^2 + t \\ p' + x = t^2 + 1 \\ x' = z \\ z' = w \\ y' = p \\ x(0) = 0 \qquad y(0) = 0 \\ z(0) = 0 \qquad p(0) = 2 \\ w(0) = 2 \end{cases}$$

This system present 3 more differential terms and so seems *more difficult*, however this formulation is numerically more suitable for the computation.

5.1 Existence of the solution

Given the general system of n ordinary differential equations in the standard form

$$\begin{cases} \boldsymbol{y}' = \boldsymbol{f}(t, \boldsymbol{y}) \\ \boldsymbol{y}(a) = \boldsymbol{y}_a \end{cases}$$

using **Peano's theorem** we can state that if the vectorial map $\boldsymbol{f} : \mathbb{R}^{n+1} \to \mathbb{R}^n$ is continuous, then a solution exists in the neighbourhood of the point $(t = a, \boldsymbol{y} = \boldsymbol{y}_a)$. This theorem provide a sufficient condition to determine if a solution exists, but it doesn't state that the solution is unique.

Example 5.2: ODE with multiple solution

Considering the ordinary differential equation in the independent variable t

$$\begin{cases} y' = \sqrt{|y|} \\ y(0) = 0 \end{cases}$$

we can see that the map $f(y) = \sqrt{|y|}$ is continuous for all $y \in \mathbb{R}$, hence for Peano's theorem a solution must exists for t *sufficiently close* to 0. In particular we can observe that the function

$$y(t) = 0$$

is a solution of the system, in fact it matches the initial condition and we have that it's derivative $y' = \frac{dy}{dt} = 0$ is equal to the function $f(y) = \sqrt{|0|} = 0$.
However this is not the lonely solution, considering in fact the function

$$y(t) = \frac{t^2}{4}\text{sign}(t) = \begin{cases} \frac{t^2}{4} & t \geq 0 \\ -\frac{t^2}{4} & t < 0 \end{cases}$$

Observing that the derivative of such function can be regarded as

$$y'(t) = \begin{cases} \frac{t}{4} & t \geq 0 \\ -\frac{t}{2} & t < 0 \end{cases} \quad \Rightarrow \quad y'(t) = \frac{|t|}{2}$$

we can also check that the provided solution solves the differential part of the system, in fact

$$\sqrt{|y(t)|} = \sqrt{\left|\frac{t^2}{4}\text{sign}(t)\right|} = \sqrt{\frac{t^2}{4}} = \frac{|t|}{2} = y'(t)$$

In general when solving system of differential equation we want to be sure that the solution exists (using as example Peano's theorem) but that is also unique. In order to do so we have to defined the **Lipschitz continuity**:

$$f : \mathbb{R}^n \to \mathbb{R} \text{ is Lip. cont.} \quad \text{if } \exists L \in \mathbb{R} \quad \text{such that} \quad \big\|f(\boldsymbol{x}) - f(\boldsymbol{y})\big\| \leq L\big\|\boldsymbol{x} - \boldsymbol{y}\big\| \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R} \tag{5.3}$$

We can so state that a system of ordinary differential equation in the standard form has one solution if the map $\boldsymbol{f} : \mathbb{R}^{n+1} \to \mathbb{R}^n$ is continuous (from Peano) and is also lipschitzian.

Example 5.3: Lipschitz continuity

Considering the problem of example 5.2 we can prove that the system has multiple solution by checking that's not Lipschitz continuous. Known that the map $f(t, y) := \sqrt{|y|}$ is continuous we can apply Lipschitz definition in the particular case when one point is the origin of the the axis:

$$\begin{aligned} \big|f(0,y) - f(0,0)\big| &\leq L\big|x - 0\big| \\ \Big|\sqrt{|y|} - \sqrt{|0|}\Big| L &\leq |y| \\ \cancel{\sqrt{|y|}} &\leq L \cancel{\sqrt{|y|}} \sqrt{|y|} \\ L &\geq \frac{1}{\sqrt{|y|}} \end{aligned}$$

Observing that for $y \to 0$ the denominator converges to zero this means that L must diverge to ∞ meaning that the function is not lipschitzian: the solution exists (Peano's theorem) but it might not unique (for Lipschitz).

5.2 Taylor expansion

The main tool used for numerical approximate solution of (system of) ordinary differential equation is the **Taylor series expansion** that's used to approximate a function $f : \mathbb{R} \to \mathbb{R}$ of class $\mathcal{C}^\infty$ as a polynomial in the neighbourhood of a specific point x_0; given h as the deviation from the point x_0 for *small* values of h we can approximate the function

$$\begin{aligned} f(x_0 + h) &\approx a_0 + a_1 h + a_2 h^2 + a_3 h^3 + \cdots + a_n h^n \\ &\approx a_0 + \sum_{k=1}^{\infty} a_k h^k \end{aligned} \tag{5.4}$$

Evaluating both members allows to obtain the first coefficient $a_0 = f(x_0)$ of the Taylor expansion, in fact

$$f(x_0) = a_0 + \sum_{k=1}^{\infty} a_k 0^k = a_0$$

By differentiation we can also obtain other coefficients

$$\begin{aligned} f'(x_0 + h) &= a_1 + \sum_{k=2}^{\infty} a_k h^{k-1} k & \xrightarrow{h=0} \quad a_1 &= f'(x_0) \\ f''(x_0 + h) &= 2a_2 + \sum_{k=3}^{\infty} a_k h^{k-2} k(k-1) & \xrightarrow{h=0} \quad 2a_2 &= f''(x_0) \end{aligned}$$

Considering that in general each coefficient can be computed as $a_k = f^{(k)}(x_0)/k!$ we can better rewrite the Taylor series expansion of equation 5.4 as

$$f(x_0 + h) \approx \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!} h^k \tag{5.5}$$

Truncation From a numerical standpoint the computation of the Taylor series is truncated to a order n and so we can use the formulation

$$f(x_0 + h) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!} h^k + R_n(h) \tag{5.6}$$

where R_n is the reminder due to the truncation of the series that can be evaluated in multiple ways:

- using Peano's formulation the more formal definition of the reminder is

  $$R_n(h) = \int_{x_0}^{x_0+h} f^{(n+1)}(s) \frac{(s-h)^n}{n!}\, ds$$

  This formulation is still complex and numerically *unusable*;

- the Lagrange reminder in the form

  $$R_n(h) = f^{(n+1)}(\zeta) \frac{h^{n+1}}{(n+1)!} \qquad \text{with } \zeta \in (x_0, x_0 + h)$$

- the *big O* notation $R_n(h) = \mathcal{O}(h^{n+1})$; in particular we denote $g(x) = \mathcal{O}(f(x))$ if

  $$\exists C \in \mathbb{R} \quad \text{such that} \quad |g(x)| \leq C|f(x)|$$

- the *small o* notation $R_n(h) = o(h^n)$; we say that $g(x) = o(f(x))$ if f is lipschitzian (equation 5.3) and we have that

  $$\lim_{x \to 0} \frac{o(f(x))}{g(x)} = 0$$

Common Taylor expansions Examples of notable series expansion on the point $x_0 = 0$ for well-known functions are the exponential, the cosine and sine:

$$\begin{aligned} e^h &= \sum_{k=0}^{\infty} \frac{h^k}{k!} = 1 + h + \frac{h^2}{2} + \frac{h^3}{3!} + \frac{h^4}{4!} + \dots \\ \cos h &= \sum_{k=0}^{\infty} -1^k \frac{h^{2k}}{2k!} = 1 - \frac{h^2}{2!} + \frac{h^4}{4!} - \frac{h^6}{6!} + \dots \\ \sin h &= \sum_{k=0}^{\infty} -1^k \frac{h^{2k+1}}{(2k+1)!} = h - \frac{h^3}{3!} + \frac{h^5}{5!} - \frac{h^7}{7!} + \dots \end{aligned} \tag{5.7}$$

Existence of the expansion Considering the definition of the lagrangian reminder in the form $f^{(m)}(\zeta)\frac{h^m}{m!}$ if we truncate the series to higher order (bigger value of m) we observe that for $h < 1$ the term $h^m/m!$ tends to zero, and so if we ensure that the m-th derivative doesn't diverge we have that the Taylor series converge to the *real* function. However this sometimes can fail: considering as example the function

$$f(x) = \begin{cases} e^{-\frac{1}{x^2}} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

it's proven that the function is *smooth* ($f \in C^\infty$) and that the k-th derivative is in the form

$$\left(e^{-\frac{1}{x^2}}\right)^{(k)} = e^{-\frac{1}{x^2}} \frac{p_1(x)}{p_2(x)} \qquad p_1, p_2 \text{ polynomials}$$

and converges to zero for $x \to 0$. By applying the definition of the Taylor expansion in $x_0 = 0$ we so have that

$$f(0+h) = \sum_{k=0}^{\infty} f^{(k)}(0)\frac{h^k}{k!} = 0$$

This expansion correctly models the left-hand side of the function f but not the right side, and so the Taylor expansion fails.

5.2.1 Multi-variable functions

Until now we have defined the Taylor expansion of function with one variable in the form $f(x)$, but such concept should be extended to function of multiple variables. Considering the simple case of $f : \mathbb{R}^2 \to \mathbb{R}$ in order to perform the Taylor series we have to ensure a continuity up to an order m for the function, meaning

$$f(x,y) \in C^m \qquad \Leftrightarrow \qquad \frac{\partial^{i+j}}{\partial x^i \partial y^j} f(x,y) \in C \quad \forall i+j \leq m \tag{5.8}$$

We have the formal definition of the Taylor series for functions of one variable (equation 5.5) and so, as idea, we can *slice* the function f passing through a point (x_0, y_0) with a direction $(x - x_0, y - y_0) = (d_x, d_y)$ using a function

$$g(t) = f(x_0 + td_x, y_0 + td_y)$$

The idea is so to compute the Taylor series of this function in the neighbourhood of $t = 0$:

$$g(t) = g(0) + g'(t)t + g''(0)\frac{t^2}{2!} + g'''(0)\frac{t^3}{3!} + \dots$$

The term $g'(t)$relates to the total derivative of $f(x_0 + td_x, y_0 + td_y)$ respect to the variable t, meaning that

$$\begin{aligned}
g'(t) &= \frac{d}{dt} f\big(x_0 + t\,d_x, y_0 + t\,d_y\big) \\
&= \frac{\partial f(\dots)}{\partial x}\frac{d(x_0 + t\,d_x)}{dt} + \frac{\partial f(\dots)}{\partial y}\frac{d(y_0 + t\,d_y)}{dt} \\
&= \frac{\partial f(\dots)}{\partial x} d_x + \frac{\partial f(\dots)}{\partial y} d_y = \frac{\partial f(\dots)}{\partial x}(x - x_0) + \frac{\partial f(\dots)}{\partial y}(y - y_0) \\
g'(0) &= \frac{\partial f(x_0, y_0)}{\partial x}(x - x_0) + \frac{\partial f(x_0, y_0)}{\partial y}(y - y_0)
\end{aligned}$$

Using a similar methodology it's possible to compute the second order derivative of g as

$$\begin{aligned}
g''(t) &= \frac{d}{dt} g'(t) = \frac{d}{dt}\left(\frac{\partial f(\dots)}{\partial x}(x - x_0) + \frac{\partial f(\dots)}{\partial y}(y - y_0)\right) \\
&= \frac{\partial^2 f(\dots)}{\partial x^2}(x - x_0)^2 + \frac{\partial^2 f(\dots)}{\partial y^2}(y - y_0)^2 + 2\frac{\partial^2 f(\dots)}{\partial x\,\partial y}(x - x_0)(y - y_0)
\end{aligned}$$

Considering as more general statement to express the Taylor series respect to a point (x_0, y_0) moving with values $h = td_x$ and $k = td_y$ the series truncated to the second order is so

$$\begin{aligned}
f(x_0 + h, y_0 + k) = & f(x_0, y_0) + \left.\frac{\partial f}{\partial x}\right|_{(x_0,y_0)} h + \left.\frac{\partial f}{\partial y}\right|_{(x_0,y_0)} k \\
& + \frac{1}{2}\left.\frac{\partial^2 f}{\partial x^2}\right|_{(x_0,y_0)} h^2 + \frac{1}{2}\left.\frac{\partial^2 f}{\partial y^2}\right|_{(x_0,y_0)} k^2 + \left.\frac{\partial^2 f}{\partial x\,\partial y}\right|_{(x_0,y_0)} hk
\end{aligned} \tag{5.9}$$

This representation can be compacted using a vectorial/matrix notation condensing (x_0, y_0) in the vector $\boldsymbol{x}_0$ and the increment $(h, k) = \boldsymbol{h}$ we can define

$$f(\boldsymbol{x}_0 + \boldsymbol{h}) = f(\boldsymbol{x}_0) + \nabla f(\boldsymbol{x}_0)\,\boldsymbol{h} + \frac{1}{2}\boldsymbol{h}^t\,\nabla^2 f(\boldsymbol{x}_0)\,\boldsymbol{h} + R_3(\|\boldsymbol{h}\|) \tag{5.10}$$

where $\nabla f(\boldsymbol{x}) = (\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n})$ is the **gradient** of the function (and is a row vector) and $\nabla^2 f(\boldsymbol{x})$ is the **hessian matrix** of f. Note that this formulation is general and is valid for any multi-variable function $f : \mathbb{R}^n \to \mathbb{R}$ assuming that's at least C^2.

Higher order Taylor expansion In order to perform Taylor expansion with order greater than 2 it's necessary to use a **tensor** notation; in particular the expansion of a function $f : \mathbb{R}^n \to \mathbb{R}$ up to the order m is described by the equation

$$f(\boldsymbol{x}_0 + \boldsymbol{h}) = \sum_{k=0}^{m} \sum_{|\alpha|=k} \partial_\alpha f(\boldsymbol{x}_0) \frac{\boldsymbol{h}^\alpha}{\alpha!} + R_m \tag{5.11}$$

where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ is the multi-index vector that consists of non-negative integers; the condition $|\alpha| = k$ based on the norm $|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_n$ choses all combination of α satisfying such relation and the multi-index variable is used for the computation following the expressions

$$\partial_\alpha := \partial_{x_1}^{\alpha_1} \partial_{x_2}^{\alpha_2} \ldots \partial_{x_n}^{\alpha_n} f(x) \qquad\qquad \boldsymbol{h}^\alpha := h_1^{\alpha_1} h_2^{\alpha_2} \ldots h_n^{\alpha_n}$$

5.3 Numerical methods

5.3.1 Taylor series based

In practise (systems of) ordinary differential equations are numerically solved by computes using algorithms some of which are based on the Taylor series expansion. Considering for the simplicity a first order ordinary differential equation in the standard form (equation 5.1)

$$\begin{cases} y' = f(x, y) \\ y(a) = y_a \end{cases}$$

If we consider $y(x)$ the solution of the ODE system, that can be expanded up to the second order with Taylor as

$$y(x + h) = y(x) + y'(x)h + \mathcal{O}(h^2) = y(x) + f(x, y)h + \mathcal{O}(h^2)$$

The most basic idea to numerically compute the solution is to subdivide the interval $[a, b]$ of integration into N pieces having each a width $h = \frac{b-a}{N}$; this discretization of the x axis determines so the sequence of point $x_k = a + hk$. With this idea in mind we can see that the yet computed Taylor expansion can be regarded as

$$y(x_{k+1}) = y(x_k) + f\big(x_k, y(x_k)\big)h + \mathcal{O}(h^2) \tag{5.12}$$

Numerical methods determines a sequence of output y_k that tends to approximate the real behaviour of the solution, hence $y_k \approx y(x_k)$. The simplest numerical method to solve the ordinary differential equation is by simply using equation 5.12 neglecting the reminder:

$$y_{k+1} = y_k + f(x_k, y_k)h \tag{5.13}$$

Error Numerical methods are approximation of the analytical solutions, hence intrinsically contains an error that should be somehow defined in order to determine *how bad* or *good* the numerical solution is. If we new define the error ϵ_k on the k-th step as the difference between the analytical and numerical solution

$$\varepsilon_{k+1} = y(x_{k+1}) - y_{k+1} = y(x_k) - y_k + \Big(f\big(x_k, y(x_k)\big) - f(x_k, y_k)\Big)h + \frac{y''(\zeta_k)}{2}h^2 \tag{5.14}$$

where the reminder $\mathcal{O}(h^2)$ as been substituted with the lagrangian one and hence $\zeta_k \in (x_k, x_{k+1})$. Considering that the function f is assumed to be lipschitzian, then it means that exists $L \in \mathbb{R}$ such that

$$\left|f\big(x_k, y(x_k)\big) - f(x_k, y_k)\right| \leq L\,|y(x_k) - y_k|$$

Considering this inequality, knowing that $y(x_k) - y_k = \epsilon_k$ and using the triangular inequality we can rewrite equation 5.14 as

$$|\varepsilon_{k+1}| = |\varepsilon_k| + hL\big|y(x_k) - y_k\big| + \frac{h^2}{2}|y''(\zeta_k)| = A|\varepsilon_k| + B \tag{5.15}$$

where $A = 1 + hL$ and $B = \frac{h^2}{2}M_2$. In particular M_2 is the constant that bounds the second derivative of y in the domain of integration, meaning

$$M_2 = \sup_{x\in[a,b]}\{y''(x)\}$$

Starting with the theoretical assumption that $\varepsilon_0 = 0$ (the initial error is null given the initial condition) and observing that $A \to 1$ and $B \to 0$, then we have that $|\varepsilon_1| \leq A0 + B = B$; the sequent error is so $|\varepsilon_2| \leq A|\varepsilon_1| + B = B(1 + A)$. Computing $|\varepsilon_3| \leq B(1 + A + A^2)$ it's possible to prove by induction that the error has a formulation

$$|\varepsilon_k| \leq \big(1 + A + A^2 + \cdots + A^{k-1}\big)B$$

The maximum error E_h of this numerical method is so determined by considering the maximum error respect to all discretization steps:

$$E_h = \max_{k=0,\ldots,N}|\varepsilon_k| \leq \max_{k=0,\ldots,N}\big(1 + A + A^2 + \cdots + A^{k-1}\big)B = \big(1 + A + A^2 + \cdots + A^{N-1}\big)B$$

Considering the geometrical series determined by $A + A^2 + \ldots$ we obtain that such sequence sums to the value $\frac{1-A^N}{1-A}$ and so the error can be considered as

$$E_h \leq \frac{A^N - 1}{A - 1}B = \frac{A^N - 1}{1 + hL - 1}\frac{h^2}{2}M_2 = \frac{A^h - 1}{L}\frac{h}{2}M_2$$

All we need now is to quantity the error related to the term A^N; considering that the Taylor series of the exponential sequence $e^x = 1 + x + \frac{x^2}{2!} + \ldots$ we have that such quantity is always greater than $1 + x$, and so knowing that $A = 1 + hL$ we have that $(1 + hL)^N \leq \left(e^{hL}\right)^n = e^{LNh}$. Observing that $Nh = b - a$ we can finally state the total error as

$$E_h \leq \frac{e^{L(b-a)}M_2}{2L}h = Ch \tag{5.16}$$

where $C \in \mathbb{R}$ is a constant, meaning that for $h \to 0$ the error presents the expected behaviour of approaching zero. By a computation point of view this method isn't that good, because in order to halve the error we have to halve also the integration step n (doubling the number of intervals N). If we would have considered other methods truncated to higher orders of derivation what we would have obtained is an error in the form

$$E_h = Ch^p$$

hence by dividing by 2 h the error would have been reduces by 2^p.

System of ODEs Considering the more general case of a system of ordinary differential equation in the standard form using the vectorial notation

$$\begin{cases}\boldsymbol{y}' = \boldsymbol{f}(t, \boldsymbol{y})\\ \boldsymbol{y}(a) = \boldsymbol{y}_a\end{cases}$$

the yet described method can be still used by expanding each component of $\boldsymbol{y}$, meaning that the numerical solution can be approximated as

$$\boldsymbol{y}_{k+1} = \boldsymbol{y}_k + h\boldsymbol{f}(x_k, \boldsymbol{y}_k)$$

and the error can be regarded as $E_h = \max\|\boldsymbol{y}(x_k) - \boldsymbol{y}_k\| \leq Ch$.

Implicit method: back-backward Euler

The numerical method described until now is the **explicit Euler integration** for determining the solution of ordinary differential equations; the formulation as provided in equation 5.13 (page 79) is computationally lightweight (because by knowing x_k, y_k at the current stage allows to automatically compute y_{k+1}), however is unstable and can quickly diverges from the analytical solution.

A way to solve such problematic is by using **implicit method** that are constructed by performing the Taylor expansion *from the left*. Considering as example the **Euler back-backward** method, the computed Taylor series is

$$y(x-h) = y(x) - hy'(x) + \frac{h^2}{2}y''(\zeta) = y(x) - hf(x,y) + \frac{h^2}{2}y''(\zeta)$$

This gives origin to the iterative numerical method defined as

$$y_{k-1} = y_k - hf(x_k, y_k) \tag{5.17}$$

where the solution of the current output y_k is implicitly defined as function of the current *position* x_k and the previous value y_{k-1}. This formulation increases the computational complexity (at each iteration a non-linear system has to be solved in order to determine the implicit solution y_k) but strongly increases the robustness of the algorithm.

Other methods based on the Taylor series

In general given the ordinary differential equation

$$\begin{cases} y' = f(x,y) \\ y(a) = y_a \end{cases}$$

in order to have a solution we assume that f is continuous and lipschitzian; given $y(x)$ the exact solution of the problem, we can apply the Taylor expansion on such result obtaining

$$y(x+h) = y(x) + hy'(x) + \frac{h^2}{2}y''(x) + \cdots + \frac{h^p}{p!}y^{(p)}(x) + \mathcal{O}(h^{p+1})$$

Knowing that $y(x)$ is the solution of the ODE, then we have that $y'(x) = f\big(x, y(x)\big)$; considering now so it's derivative we have that

$$\begin{aligned} y''(x) &= \frac{d}{dx}y'(x) = \frac{d}{dx}f\big(x,y(x)\big) = \frac{\partial f}{\partial x}\big(x,y(x)\big) + \frac{\partial f}{\partial y}\big(x,y(x)\big)y'(x) \\ &= \frac{\partial f}{\partial x}\big(x,y(x)\big) + \frac{\partial f}{\partial y}\big(x,y(x)\big)f\big(x,y(x)\big) \end{aligned}$$

This allows to rewrite the Taylor expansion as

$$y(x+h) = y(x) + hy'(x) + \frac{h^2}{2}\left(\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}f\right) + \cdots + \frac{h^p}{p!}y^{(p)}(x) + \mathcal{O}(h^{p+1})$$

The previously described explicit Euler method was determined by neglecting the terms with order higher then h, but in this case we have the possibility to express also y'' as function of f and x, y increasing hence the numerical accuracy. The numerical method is so

$$y_{k+1} = y_k + hf(x_k, y_k) + \frac{h}{2}\left(\frac{\partial f(x_k,y_k)}{\partial x} + \frac{\partial f(x_k,y_k)}{\partial y}f(x_k,y_k)\right) \tag{5.18}$$

where so in this case we dropped an error in the form $\mathcal{O}(h^3)$. Increasing the order of the numerical method increases the solution but requires the symbolical computation of the derivatives $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$; we can

carry on the process to explicitly determine all the derivative $y^{(p)}$ (up to a certain order) as function of x, y and partial derivatives of f. Considering as example the third derivative of y we see that

$$\begin{aligned} y'''(x) &= \frac{d}{dx}y''(x) \\ &= \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial x\,\partial y}y' + \frac{\partial^2 f}{\partial x\,\partial y}f + \frac{\partial^2 f}{\partial y^2}yf + \frac{\partial f}{\partial y}\left(\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}y'\right) \end{aligned}$$

We can see that numerical method based on the Taylor series are conceptually easy but requires *a lot* of symbolical computation in order to express the derivatives as function of the known variables.

Example 5.4: explicit Euler scheme of the second order

Given the ODE

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

to express the 2^{nd} order Euler scheme we need to compute the partial derivatives

$$\frac{\partial f}{\partial x} = y + 1 \qquad\qquad \frac{\partial f}{\partial y} = x$$

and hence using equation 5.18 we obtain the method

$$y_{k+1} = y_k + h\big(x_k y_k + x_k\big) + \frac{h^2}{2}\Big[y_k + 1 + x_n\big(x_n y_n + x_n\big)\Big]$$

Higher order methods for this problem can be implemented explicitly computing the derivatives

$$\begin{aligned} y'(x) &= xy(x) + x \\ y''(x) &= y(x) + xy'(x) + 1 \\ y'''(x) &= y'(x) + y'(x) + xy''(x) = 2y'(x) + xy''(x) \\ y^{(4)}(x) &= 2y''(x) + y''(x) + xy'''(x) = 3y''(x) + xy'''(x) \\ &\vdots \end{aligned}$$

The idea is so to determine the numerical method as

$$y_{k+1} = y_k + hy'_k + \frac{h^2}{2}y''_k + \frac{h^3}{6}y'''_k + \frac{h^4}{24}y^{(4)}_k + \dots$$

where

$$\begin{aligned} y'_k = x_k y_k + x_k \quad &\to \quad y''_k = y_k + x_k y'_k + 1 \\ \to \quad y'''_k = 2y'_k + x_k y''_k \quad &\to \quad y^{(4)}_k = 3y''_k + xy'''_k \quad \to \quad \dots \end{aligned}$$

Economic Taylor scheme Recalling the higher order method shown in the previous example, a way to simplify the notation on the numerical method based on the Taylor series we consider the scheme

$$y_{k+1} = y_k + hy'_k + \frac{h^2}{2}y''_k + \dots + \frac{h^p}{p!}y^{(p)}_k$$

where the numerical derivatives are recursively defined considering that $y'_k = f(x_k, y_k)$:

$$\begin{aligned}
y'_k = D_1\big(x, y(x)\big) &= f(x,y) \\
y''_k = D_2\big(x, y(x), y'(x)\big) &= \frac{\partial D_1}{\partial x}(x,y) + \frac{\partial D_1}{\partial y}(x,y)\, y'(x,y) \\
y'''_k = D_3\big(x, y(x), y'(x), y''(x)\big) &= \frac{\partial D_2}{\partial x}(x,y) + \frac{\partial D_2}{\partial y} y' + \frac{\partial D_2}{\partial y'} y'' \\
&\vdots \\
y^{(p)}_k = D_p\big(x, y(x), y'(x), \dots, y^{(p-1)}(x)\big) &= \frac{\partial D_{p-1}}{\partial x} + \frac{\partial D_{p-1}}{\partial y} + \cdots + \frac{\partial D_{p-1}}{\partial y^{(p-2)}} y^{(p-1)}
\end{aligned}$$

5.3.2 Runge-Kutta

Considering the statements seen for numerical methods derived from the Taylor series expansion, we can see that the function f can be expanded by parameters α, β as

$$f(x+\alpha, y+\beta) = f(x,y) + \frac{\partial f}{\partial x}\alpha + \frac{\partial f}{\partial y}\beta + \frac{1}{2}\frac{\partial^2 f}{\partial x^2}\alpha^2 + \frac{\partial^2 f}{\partial x\, \partial y}\alpha\beta + \frac{1}{2}\frac{\partial^2 f}{\partial y^2}\beta^2 + \mathcal{O}\Big(\sqrt{\alpha^2+\beta^2}^{\,3}\Big)$$

The idea that originates the Runge-Kutta method is so to combine many evaluation of $f(x+\alpha, y+\beta)$ in order to match the results provided by the Taylor series; in general this match determines an error that however should be at maximum comparable to the order $\mathcal{O}(x^n)$ required. Considering the expansion previously discussed

$$y(x+h) = y(x) + hf\big(x, y(x)\big) + \frac{h^2}{2}\left(\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} f\right) + \mathcal{O}(h^3)$$

we can consider the term

$$(i): \qquad h\, f\big(x, y(x)\big) + \frac{h^2}{2}\left(\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} f\big(x, y(x)\big)\right)$$

and it's Taylor expansion

$$\begin{aligned}
h\omega f\big(x, y(x)\big) + h\gamma f\big(x+\alpha h, y(x)+\beta h\big) &= h\omega f + h\gamma\left(f + \frac{\partial f}{\partial x}\alpha h + \frac{\partial f}{\partial y}\beta h + \mathcal{O}(h^2)\right) \\
(ii): \qquad &= h\omega f + h\gamma f + h^2\gamma\alpha\frac{\partial f}{\partial x} + h^2\gamma\beta\frac{\partial f}{\partial y} + \mathcal{O}(h^3)
\end{aligned}$$

By now subtracting (ii) to (i) we determine

$$hf(1-\omega-\gamma) + h^2\frac{\partial f}{\partial x}\left(\frac{1}{2} - \gamma\alpha\right) + h^2\frac{\partial f}{\partial y}\left(\frac{f}{2} - \beta\gamma\right) + \mathcal{O}(h^3)$$

In order to reduce the error with a threshold $\mathcal{O}(h^3)$ we have to set to zero all the multiplicative terms of f (and it's derivatives) determining the following non-linear system:

$$\begin{cases} 1-\omega-\gamma = 0 \\ \frac{1}{2} - \gamma\alpha = 0 \\ \frac{f}{2} - \beta\gamma = 0 \end{cases}$$

If we determine parameters $\omega, \gamma, \alpha, \beta$ that satisfy such conditions we have that the expansion

$$y(x+h) = y(x) + \omega\, f\big(x, y(x)\big) + \gamma\, f\big(x+\alpha h, y(x)+\beta h\big) + \mathcal{O}(h^3)$$

matches the result obtained with Taylor; the system has 3 equation but the unknowns are 4, having so the a parametric solution in the form $\omega = 1-\gamma$, $\alpha = \frac{1}{2\gamma}$ and $\beta = \frac{f}{2\gamma}$ is always satisfied. Substituting this in the original expression we have that all the 2^{nd} order numerical method that do not use *explicitly* the partial derivatives of $f(x,y)$ are

$$y(x+h) = y(x) + h(1-\gamma)f(x,y(x)) + h\gamma f\left(x + \frac{h}{2\gamma}, y(x) + \frac{h}{2\gamma}f(x,y(x))\right) + \mathcal{O}(h^3)$$

determining the numerical method

$$y_{k+1} = y_k + h(1-\gamma)f(x_k,y_k) + h\gamma f\left(x_k + \frac{h}{2\gamma}, y_k + \frac{h}{2\gamma}f(x_k,y_k)\right) \tag{5.19}$$

Definition The idea of the **Runge-Kutta** method is use a combination of *displacements* in order to match *as much as possible* the Taylor expansion of the exact solution; in particular the numerical steps are written as

$$\boldsymbol{y}_{k+1} = \boldsymbol{y}_k + \sum_{i=1}^{s} b_j \boldsymbol{k}_j \tag{5.20}$$

where the s vectors $\boldsymbol{k}_j$ (where s is the order of the Runge-Kutta method) are obtained as

$$\begin{cases} \boldsymbol{k}_1 = h\,\boldsymbol{f}\left(x_k + c_1 h, y_k + \sum_{j=1}^{s} A_{1j}\boldsymbol{k}_j\right) \\ \boldsymbol{k}_2 = h\,\boldsymbol{f}\left(x_k + c_2 h, y_k + \sum_{j=1}^{s} A_{2j}\boldsymbol{k}_j\right) \\ \quad\vdots \\ \boldsymbol{k}_s = h\,\boldsymbol{f}\left(x_k + c_s h, y_k + \sum_{j=1}^{s} A_{sj}\boldsymbol{k}_j\right) \end{cases} \tag{5.21}$$

where the coefficients c_j, b_j, A_{ij} are computed in such a way that $y_{k+1} - y(x_{k+1}) = \mathcal{O}(h^p)$ (so the error between the computed value and the theoretical solution) where p is as large as possible. Such values are already tabled in the ***Runge-Kutta tableaus*** represented as

$$\begin{array}{c|c} \boldsymbol{c} & A \\ \hline & \boldsymbol{b}^t \end{array} \tag{5.22}$$

where $\boldsymbol{c} = (c_1, \dots, c_s)$, $\boldsymbol{b} = (b_1, \dots, b_s)$ (with $\boldsymbol{c}, \boldsymbol{b} \in \mathbb{R}^s$) and $A \in \mathbb{R}^{s\times s}$.

Runge-Kutta of order 4 A tableau for the Runge-Kutta method of order 4 is defined as

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{3} & \frac{1}{3} & & & \\ \frac{2}{3} & -\frac{1}{3} & -1 & & \\ 1 & 1 & -1 & 1 & \\ \hline & \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{array}$$

where the non-represented terms are zeros. Recalling equation 5.20 as the definition of the Runge-Kutta, we have that the numerical method derived from that is

$$y_{k+1} = y_j + \frac{1}{8}k_1 + \frac{3}{8}k_2 + \frac{3}{8}k_3 + \frac{1}{8}k_4$$

where

$$\begin{cases} k_1 &= h\,f(x_k, y_k) \\ k_2 &= h\,f\left(x_k + \frac{1}{3}h, y_k + \frac{1}{3}k_1\right) \\ k_3 &= h\,f\left(x_k + \frac{2}{3}h, y_k - \frac{1}{3}k_1 - k_2\right) \\ k_4 &= h\,f\left(x_k + h, y_k + k_1 - k_2 + k_3\right) \end{cases}$$

In this case the method is **explicit**: in fact we can sequentially compute k_1 (that depends on the knowns h, x_k, y_k) and sequentially k_2, then k_3 and lastly k_4.

Euler methods If we consider the *strange* tableau

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$$

what we obtain is the explicit Euler method, in fact having $k_1 = h\,f(x_k, y_k)$ determines the method

$$y_{k+1} = y_k + 1k_1 = y_k + hf(x_k, y_k)$$

and perfectly matches the definition provided in equation 5.13 at page 79. Considering now instead the tableau

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$$

it determines an **implicit** method: we in fact have that $k_1 = h\,f(x_k + h, y_k + k_1)$ where the solution is implicit in k_1; considering that such relation can be regarded as $k_1 = hf(x_{k+1}, y_{k+1})$ what we determine is the implicit Euler method (equation 5.17, page 81)

$$y_{k+1} = y_k + h\,f\big(x_{k+1}, y_{k+1}\big)$$

If we in general have that all the elements A_{ij} with $j \geq j$ are all zeros, then the Runge-Kutta scheme is explicit meaning that all the coefficients $\boldsymbol{k}_1, \boldsymbol{k}_2, \ldots, \boldsymbol{k}_s$ can be computed consecutively. *Graphically* this means that the matrix A must have non-zero terms only below its principal matrix not included (it means that A_{ii} must always be zero).

One step methods

Usually we refer to ***one step methods*** the ones that allow to explicitly compute the next step as function of the current step in the form

$$y_{k+1} = \phi(x_k, y_h, h)$$

An example is the explicit Euler method (equation 5.13, page 79) that's characterized by the function $\phi(x_k, y_k, h) = y + hf(x_k, y_k)$. Also implicit methods can be one step; considering the implicit Euler (equation 5.17, page 81) it can be considered as

$$y_{k+1} = y_k + K_1 \qquad \text{with } K_1 = h\,f\big(x_{k+1}, y_k + K_1\big)$$

the value K_1 is formally a function of the parameters x_k, y_k, h, hence

$$G\big(K_1, x_k, y_k, h\big) = K - h\,f\big(x_k + hy_k + K\big) \qquad \Rightarrow \qquad \phi(x, y, h) = y + K(x, y, h)$$

In general all Runge-Kutta methods (both explicit and implicit) can be formally written in the form $y_{k+1} = \phi(x_k, y_k, h)$ and so are one step methods.

Error propagation

Known that each Runge-Kutta is a one-step method, then we can express the **local truncation error** $\tau_k(h)$ as the difference between the theoretical computed value and the numerical result obtained:

$$y\big(x_{k+1}\big) = \phi\big(x_k, y(x_k), h\big) + \tau_k(h) \qquad \Rightarrow \qquad \tau(h) = y(x+h) - \phi\big(x, y(x), h\big)$$

Considering the **error** $\varepsilon_k = y\big(x_k\big) - y_k$ as the difference between the analytical solution and the numerical approximated one, we can regard the two cases as

$$\begin{aligned} (i)&: \qquad y_{k+1} = \phi(x_k, y_k, h) \\ (ii)&: \qquad y\big(x_{k+1}\big) = \phi\big(x_k, y(x_k), h\big) + \tau_k(h) \end{aligned}$$

performing the difference $(i) - (ii)$ what we obtain is

$$\varepsilon_{k+1} = \Big(\phi\big(x_k, y(x_k), h\big) - \phi(x_k, y_k, h)\Big) + \tau_k(h)$$

Considering the simple case of the explicit Euler method characterized by a function $\phi(x, y, h) = y + h\, f(x, y)$, the difference in the parenthesis is in the form $\phi(x, z, h) - \phi(x, y, h) = z - y + h\big(f(x, z) - f(x, y)\big)$; computing it's absolute value we have and considering that f is lipschitzian we have

$$\big|\phi(x, z, h) - \phi(x, y, h)\big| \leq |z - y| + h\big|f(x, z) - f(x, y)\big| \leq (1 + hL)|z - y|$$

We can so rewrite the magnitude of the error as

$$|\varepsilon_{k+1}| \leq (1 + hL)\big|y(x_k) - y_k\big| + |\tau_k(h)| \leq (1 + hL)|\varepsilon_k| + |\tau_k(h)|$$

MIN 7.08

Chapter 6

Introduction to Algebraic Differential Equations

The **algebraic differential equations** **DAEs** can be regarded as a system of ordinary differential equations combined with *general* algebraic equations; as example a DAE system is

$$\begin{cases} y' = f(x,y) \\ y(a) = y_a \\ f(x,y) = 0 \\ x^2 - y = 3 \end{cases}$$

In general for ODEs and algebraic equations a lot of numerical methods have been implemented with consolidated theory regarding existence, stability... The problem is that when combining such theories in the algebraic differential equations, the numerical results that we might wanna retrieve are a *nightmare* to compute.

DAE with an example: simple pendulum Considering the simple pendulum of a mass m fixed by a bar of length l to a pivot point; considering such center as the origin of a reference frame, the coordinates of the mass can be described as function of using the minimal number of coordinates (associated in this case to lagrangian coordinate θ as the angle between the bar and the vertical line) as

$$x = l\sin\theta \qquad\qquad y = -l\cos\theta$$

The idea is that this coordinates satisfy the constraint $x^2 + y^2 = l^2$, meaning that the mass m can move only on the circle of radius l. Taking the velocities

$$\dot{x} = \frac{dx}{dt} = l\cos\theta\,\dot{\theta} \qquad\qquad \dot{y} = \frac{dy}{dt} = l\sin\theta\,\dot{\theta}$$

Computing the kinematic and potentials energy as

$$T = \frac{m}{2}(\dot{x}^2 + \dot{y}^2) = \frac{m}{2}l^2\dot{\theta}^2 \qquad\qquad V = mgy = -mgl\cos\theta$$

With this we can build the lagrangian $\mathcal{L} = T - V = \frac{m}{2}l^2\dot{\theta}^2 + mgl\cos\theta$ and using the Euler-Lagrange equation (following the minimal action principle) the differential equation describing the motion is

$$\frac{d}{dt}\frac{\partial\mathcal{L}}{\partial\dot{\theta}} - \frac{\partial\mathcal{L}}{\partial\theta} = ml^2\ddot{\theta} + mgl\sin\theta = l\ddot{\theta} + g\sin\theta = 0$$

where the ordinary differential equation, in order to be solved/integrated, requires the initial conditions $\theta(0) = \theta_0$ and $\dot{\theta}(0) = \dot{\theta}_0$. Introducing $\omega = \dot{\theta}$ we can reduce the system of ODEs to the first order

that can be numerically solved:

$$\begin{cases} \dot{\theta} = \omega \\ l\dot{\omega} + g\sin\theta = 0 \\ \theta(0) = \theta_0, \quad \omega(0) = \omega_0 = \dot{\theta}_0 \end{cases}$$

Observe that we obtained the solution as an ordinary differential equation because we found the *minimal* set of coordinates which describes the system.

As alternative approach we could have used simpler independent ordinary differential equations and add some constraints; considering the independent mass m described by the point (x, y) in the plane and constrained to move on a circle of radius l, it's kinetic and potential energies are still $T = \frac{m}{2}(\dot{x}^2 + \dot{y}^2)$ and $V = mgy$ (note that no transformation in terms of θ has been applied), then the lagrangian is

$$\mathcal{L} = T - V = \frac{m}{2}(\dot{x}^2 + \dot{y}^2) - mgy$$

The constraint is described by the equation $\phi(x, y) = x^2 + y^2 - l^2 = 0$. Adding to the constraint the least action principle stating that the functional $\mathcal{A} = \int_{t_0}^{t_1} \mathcal{L}(x, y, \dot{x}, \dot{y}, t)\, dt$ we can find the *stationary point* of the action $\mathcal{A}$ that's subject to $\phi(x, y) = 0$. Expanding the definition

$$\int_{t_0}^{t_1} \mathcal{L}(x, y, \dot{x}, \dot{y}, t) - \lambda\phi(x, y)\, dt$$

we can build the hamiltonian $\mathcal{H} = \mathcal{L} - \lambda\phi$ that, after the first variation, determines the system

$$\begin{cases} \frac{d}{dt}\frac{\partial \mathcal{H}}{\partial \dot{x}} - \frac{\partial \mathcal{H}}{\partial x} & = m\ddot{x} + \lambda x = 0 \\ \frac{d}{dt}\frac{\partial \mathcal{H}}{\partial \dot{y}} - \frac{\partial \mathcal{H}}{\partial y} & = m\ddot{y} + \lambda y = -mg \\ \phi(x, y) & = x^2 + y^2 - l^2 = 0 \end{cases}$$

that is a **differential algebraic equation**; introducing $\dot{x} = u$ and $\dot{y} = v$ we can simplify to a differential algebraic equation

$$\begin{cases} m\dot{u} + \lambda x = 0 \\ m\dot{v} + \lambda y = -mg \\ \dot{x} = u, \qquad \dot{y} = v \\ x^2 + y^2 - l^2 = 0 \end{cases} \tag{6.1}$$

Introduction to numerical methods for DAEs Considering the example of the pendulum, we can rewrite the differential equations as

$$\begin{pmatrix} \dot{x} \\ \dot{y} \\ \dot{u} \\ \dot{v} \\ 0 \end{pmatrix} = \begin{pmatrix} u \\ v \\ -\lambda x/m \\ -\lambda y/m - g \\ x^2 + y^2 - l^2 \end{pmatrix}$$

Defining $\boldsymbol{z} = (x, y, u, b, \lambda)$, such relation is similar to the form $\dot{\boldsymbol{z}} = \boldsymbol{F}(t, \boldsymbol{z})$. The main idea is in fact to **transform DAEs to ODEs**; note in fact that the last equation is the lonely one that's not already a ordinary differential equation: deriving it in time determines

$$\frac{d}{dt}(x^2 + y^2 - l^2) = 2x\dot{x} + 2y\dot{y} = 2xu + 2yv$$

but stull we observe that the variable λ is missing in the equation. Deriving one more time respect to t we obtain

$$\begin{aligned} \frac{d}{dt}(2xu + 2yv) &= 2\dot{x}u + 2x\dot{u} + 2\dot{y}v + 2y\dot{v} = 2u^2 + 2v^2 - 2x\frac{\lambda x}{m} - 2y\left(\frac{\lambda y}{m} + g\right) \\ &= 2(u^2 + v^2) - \frac{2\lambda}{m}(x^2 + y^2) - 2yg \end{aligned} \tag{6.2}$$

By substituting the different known relations for $\dot{x}, \dot{y}, \dot{u}, \dot{v}$ we so obtain a derivative that's function of λ, but not of $\dot{\lambda}$. Deriving one more time respect to the variable t

$$\begin{aligned}\frac{d}{dt}(6.2) &= 4\left(u\dot{u} + v\dot{v}\right) - \frac{4\lambda}{m}\left(x\dot{x} + y\dot{y}\right) - 2\dot{y}g - \frac{2}{m}\dot{\lambda}\left(x^2 + y^2\right) \\ &= 4\left(-u\frac{\lambda x}{m} - v\frac{\lambda y}{m} - vg\right) - \frac{4\lambda}{m}\left(xu + yv\right) - 2vg - \frac{2}{m}\dot{\lambda}\left(x^2 + y^2\right) = 0\end{aligned}$$

Solving for $\dot{\lambda}$ so gives

$$\dot{\lambda} = \frac{-4\lambda\left(xy + yv\right) - 3vmg}{x^2 + y^2}$$

We can so rewrite the differentia algebraic system in equation 6.1 as a system of ODE only as

$$\begin{cases} \dot{x} = u \\ \dot{y} = v \\ \dot{u} = -\frac{l}{m}x \\ \dot{v} = -\frac{\lambda}{m}y - g \\ \dot{\lambda} = \frac{-4\lambda\left(xy+yv\right)-3vmg}{x^2+y^2} \end{cases}$$

With such definition we can use numerical methods to solve the form $\dot{\boldsymbol{z}} = \boldsymbol{F}(t, \boldsymbol{z})$. This formulations however introduces some problems:

- the initial condition on λ is not set. This problem can be overcome considering that given x, the coordinate y is constrained by $x^2 + y^2 = l^2$. If we moreover know $\dot{x} = u$ then using the derivative of the constraint $2xy + 2yv = 0$, then also $v = \dot{y}$ is constrained. Using the second derivative of the constraint (equation 6.2) we can finally solve for λ_0. We see that the initial conditions must satisfy the *original* constraints and the *hidden ones* determined by the derivatives of the algebraic equations.
- another problem is that if we considered a constraint in the form

  $$\phi(x, y) = x^2 + y^2 - l^2 + a + bt + ct^2 = 0$$

  in order to obtain the ODE equivalent system we have to derive ϕ three times over time resulting in the cancellations of the polynomial terms $a + bt + ct^2$ (we in fact would have obtained the same ODE system).

6.1 Linear differential algebraic equations

Starting from the simplest cases of study, a generic differential algebraic equation can be written as

$$\begin{cases} \boldsymbol{F}(t, \boldsymbol{y}, \boldsymbol{y}') = 0 \\ \boldsymbol{y}(a) = \boldsymbol{y}_a \end{cases}$$

with $\boldsymbol{y}(t) \in \mathbb{R}^n$. We can say that the map $\boldsymbol{F}$ is **linear** in y if it can be expressed as linear combination of $\boldsymbol{y}'$ in the form

$$\boldsymbol{F}(t, \boldsymbol{y}, \boldsymbol{y}') = \mathbf{E}(t, \boldsymbol{y})\boldsymbol{y}' + \boldsymbol{G}(t, \boldsymbol{y}) \tag{6.3}$$

Moreover the map $\boldsymbol{F}$ is linear in both $\boldsymbol{y}$ and $\boldsymbol{y}'$ if it can be regarded as

$$\boldsymbol{F}(t, \boldsymbol{y}, \boldsymbol{y}') = \mathbf{E}(t)\boldsymbol{y}' + \mathbf{A}(t)y - \boldsymbol{C}(t) \tag{6.4}$$

where in general $\mathbf{E}, \mathbf{A}$ are matrices that can sometimes be singular. The map $\boldsymbol{F}$ is said **linear with constant coefficients** if it happens that the matrices $\mathbf{E}, \mathbf{A}$ are time independent.

Example 6.1: linear DAEs

An example of linear differential algebraic equation is the one that can be described as

$$\begin{bmatrix} 1 & t \\ t^2 & 2 \end{bmatrix} \begin{pmatrix} y_1' \\ y_2' \end{pmatrix} + \begin{bmatrix} \sin t & \cos t \\ t^2 & 1 \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} - \begin{pmatrix} e^t \\ 1+t \end{pmatrix}$$

Observe that if the matrix $\boldsymbol{E}$ in non singular, expression 6.3 corresponds to a *simple* ordinary differential equation: it can be in fact rewritten as

$$\boldsymbol{y}' = -\mathbf{E}^{-1}(t, \boldsymbol{y})\boldsymbol{G}(t, \boldsymbol{y}) = \boldsymbol{f}(t, \boldsymbol{y}) \tag{6.5}$$

For the moment we can assume that if $\mathbf{E}$ is singular the system is not an ODE. In this first part we will focus on linear differential algebraic equations in order to exploit linear algebra tools to ease the calculations.

Jordan normal form As a recall from the linear algebra, given a matrix $\mathbf{B} \in \mathbb{R}^{n \times n}$ there exists always a non-singular matrix $\mathbf{T} \in \mathbb{R}^{n \times n}$ such that

$$\mathbf{T}^{-1}\mathbf{B}\mathbf{T} = \mathbf{J}$$

where $\mathbf{J}$ is the **Jordan matrix form** defined as

$$\mathbf{J} = \begin{bmatrix} J_1 & & & 0 \\ & J_2 & & \\ & & \ddots & \\ 0 & & & J_m \end{bmatrix} \qquad \text{where } J_k = \begin{bmatrix} \lambda_k & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda_k \end{bmatrix} \tag{6.6}$$

Regular pencil Given the matrices $\mathbf{B}, \mathbf{C} \in \mathbb{R}^{n \times n}$, the couple $(\mathbf{B}, \mathbf{C})$ is a **regular pencil** if

$$f(\lambda) = \det\left(\mathbf{B} - \lambda\mathbf{C}\right) \neq 0$$

is not identically null, or equivalently if there exists a λ such that $f(\lambda) \neq 0$. Considering as example the matrices

$$\mathbf{B} = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \qquad \mathbf{C} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

is not a regular pencil, in fact the polynomial $f(\lambda) = \det(\mathbf{B} - \lambda\mathbf{C}) = \det \begin{bmatrix} 1-\lambda & 1 \\ 0 & 0 \end{bmatrix} = 0$ always evaluates to zero.

Nilpotent matrix A matrix $\mathbf{B} \in \mathbb{R}^{n \times n}$ is **nilpotent** of order p if

$$\mathbf{B}^p = 0 \quad \text{and} \quad \mathbf{B}^j \neq 0 \; \forall j < p \tag{6.7}$$

where $\mathbf{B}^p$ is the product $\mathbf{B}\mathbf{B}\ldots\mathbf{B}$ p times. Considering as example

$$\mathbf{B} = \begin{bmatrix} 0 & 1 & 2 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{bmatrix} \qquad \mathbf{B}^2 = \begin{bmatrix} 0 & 0 & 3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad \mathbf{B}^3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

we have that such matrix $\mathbf{B}$ is nilpotent of order 3. Observe that if the matrix $\mathbf{B}$ is non-singular, than it can't be nilpotent.

Kronecker normal form If we consider two regular pencil matrices $(\mathbf{B}, \mathbf{C}) \in \mathbb{R}^{n\times n}$, then there exists two non-singular matrices $\mathbf{P}, \mathbf{Q} \in \mathbb{R}^{n\times n}$ such that

$$\mathbf{PBQ} = \begin{bmatrix} \mathbf{N} & 0 \\ 0 & \mathbf{I} \end{bmatrix} \qquad \text{and} \qquad \mathbf{PCQ} = \begin{bmatrix} \mathbf{I} & 0 \\ 0 & \mathbf{J} \end{bmatrix} \tag{6.8}$$

where $\mathbf{N}$ is a nilpotent matrix, $\mathbf{I}$ is the identity matrix and $\mathbf{J}$ is a Jordan normal form matrix. Considering that the *blocks* $\mathbf{N}, \mathbf{J}$ can be empty, as extreme cases we have $\mathbf{PBQ} = \mathbf{I}$, $\mathbf{PCQ} = \mathbf{J}$, $\mathbf{PBQ} = \mathbf{N}$ and $\mathbf{PCQ} = \mathbf{I}$.

6.1.1 Usage of the Kronecker normal form

To ease the computation of linear differential algebraic equation, we can use the Kronecker normal form assuming that the couple of matrices $(\mathbf{E}, \mathbf{A})$ (equation 6.4) are a regular pencil (in order not to have an *inconsistent* DAE).

Example 6.2: inconsistent DAE

Using the matrices defined used in the theory of regular pencil, we can build a DAE system of the form

$$\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \begin{pmatrix} y_1' \\ y_2' \end{pmatrix} + \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} t \\ 1 \end{pmatrix}$$

the associated system is

$$\begin{cases} y_1' + y_2' + y_2 = t \\ 0 = 1 \end{cases}$$

that's inconsistent.

With such assumption we can compute the Kronecker normal form $\mathbf{PEQ} = \begin{bmatrix} \mathbf{N} & \\ & \mathbf{I} \end{bmatrix}$ and $\mathbf{PAQ} = \begin{bmatrix} \mathbf{I} & \\ & \mathbf{J} \end{bmatrix}$; premultiplying so equation 6.4 by $\mathbf{P}$ results in $\mathbf{PE}\boldsymbol{y}' + \mathbf{PA}\boldsymbol{y} = \mathbf{P}\boldsymbol{C}$. Performing the change of variable $\mathbf{Q}\boldsymbol{z} = \boldsymbol{y}$ (hence $\boldsymbol{z} = \mathbf{Q}^{-1}$) and observing that $\mathbf{Q}\boldsymbol{z}' = \boldsymbol{y}'$ we obtain the expression $\mathbf{PEQ}\boldsymbol{z}' + \mathbf{PAQ}\boldsymbol{z} = \mathbf{P}\boldsymbol{C}$ on top of which we can apply the Kronecker normal form:

$$\begin{bmatrix} \mathbf{N} & \\ & \mathbf{I} \end{bmatrix} \boldsymbol{z}' + \begin{bmatrix} \mathbf{I} & \\ & \mathbf{J} \end{bmatrix} \boldsymbol{z} = \mathbf{P}\boldsymbol{C} \tag{6.9}$$

Splitting both vectors $\boldsymbol{z} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $\mathbf{P}\boldsymbol{C} = (\boldsymbol{d}, \boldsymbol{e})$ we can rewrite this expression as

$$\begin{bmatrix} \mathbf{N} & \\ & \mathbf{I} \end{bmatrix} \begin{pmatrix} \boldsymbol{\alpha}' \\ \boldsymbol{\beta}' \end{pmatrix} + \begin{bmatrix} \mathbf{I} & \\ & \mathbf{J} \end{bmatrix} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} = \begin{pmatrix} \boldsymbol{d} \\ \boldsymbol{e} \end{pmatrix}$$

The associated linear system representation is

$$\begin{cases} i) & \mathbf{N}\boldsymbol{\alpha}' + \boldsymbol{\alpha} = \boldsymbol{d}(t) \\ ii) & \boldsymbol{\beta}' + \mathbf{J}\boldsymbol{\beta} = \boldsymbol{e} \end{cases}$$

$ii)$ represent a *standard* system of ordinary differential equations; the term $i)$ is instead more complex but we can invert the relation to obtain $\boldsymbol{\alpha} = \boldsymbol{d} - \mathbf{N}\boldsymbol{\alpha}'$. Observing that the k-th derivative in time of this expression evaluates to $\boldsymbol{\alpha}^{(k)} = \boldsymbol{d}^{(k)} - \mathbf{N}\boldsymbol{\alpha}^{k+1}$, we can substitute the derivatives $\boldsymbol{\alpha}^{(k)}$ determining

$$\begin{aligned} \boldsymbol{\alpha} &= \boldsymbol{d} - \mathbf{N}\boldsymbol{\alpha}' = \boldsymbol{d} - \mathbf{N}(\boldsymbol{d}' - \mathbf{N}\boldsymbol{\alpha}'') = \boldsymbol{d} - \mathbf{N}\boldsymbol{d}' + \mathbf{N}^2\boldsymbol{\alpha}'' = \boldsymbol{d} - \mathbf{N}\boldsymbol{d}' + \mathbf{N}^2(\dots) = \dots \\ &= \boldsymbol{d} - \mathbf{N}\boldsymbol{d}' + \mathbf{N}^2\boldsymbol{d}'' - \mathbf{N}^3\boldsymbol{d}''' + \mathbf{N}^4\boldsymbol{d}^{(4)} - \mathbf{N}^5\boldsymbol{d}^{(5)} + \dots \end{aligned}$$

This series is infinite, however being $\mathbf{N}$ a nilpotent matrix of order p we have only that the first p terms remains and so

$$\boldsymbol{\alpha} = \boldsymbol{d} - \mathbf{N}\boldsymbol{d}' + \mathbf{N}^2\boldsymbol{d}'' + \cdots + (-1)^{p-1}\mathbf{N}^{p-1}\boldsymbol{d}^{(p-1)} + \cancel{(-1)^p\mathbf{N}^p\boldsymbol{d}^{(p)}} = \sum_{j=0}^{p-1}(-1)^j\mathbf{N}^j\boldsymbol{d}^{(j)} \tag{6.10}$$

With this relation we determined $\boldsymbol{\alpha}$ without using the initial conditions (as was required for the DEA solution using the conversion to ODE), depends only on $\boldsymbol{d}$ (and it's derivatives up to the $p-1$ order). Also observe that $ii)$ is a *regular* ODE, hence the initial values $\boldsymbol{\beta}(0)$ must be specified.

The order of the nilpotency of the matrix $\mathbf{N}$ is the so called **index** of the differential algebraic equation (for the linear ones) and is a sort of *measure* of the *difficulty* of solving numerically the DAE. In particular if $p = 0$ then what we have is a system of ordinary differential equation while if $p = n$ we have a set of only algebraic equations.

6.1.2 LU decomposition and Jacobi modification

In general the Kronecker normal form is *hard* to compute; computationally we can use the *simpler* **LU decomposition** in order to reduce the index of a differential algebraic equation, or better transform the DAE in an ordinary differential equation.

Recall on the LU decomposition Considering $\mathbf{A} \in \mathbb{R}^{n\times n}$ a square matrix, then there exists 2 permutation matrices $\mathbf{P}, \mathbf{Q}$ such that

$$\mathbf{PAQ} = \mathbf{LU} \tag{6.11}$$

where $\mathbf{L}$ is a **lower triangular matrix** and $\mathbf{U}$ is an **upper** triangular one (in particular if the matrix $\mathbf{A}$ is singular only the first $m < n$ rows of $\mathbf{U}$ are non-zero and are still triangular, having a *trapezoidal shape*).

In the case that $\mathbf{A}$ is non-singular, then the algorithm can be reduced to the form $\mathbf{PA} = \mathbf{LU}$; alternatively we can use the **Echelon form** determined as $\mathbf{PA} = \mathbf{LUQ}^T = \mathbf{L\tilde{U}}$.

A **permutation matrix** S_{ij} are used to exchange the i-th and j-th row/column of a matrix A; in particular $S_{ij}A = \tilde{A}$ results in a swapping of the rows while $AS_{ij} = \tilde{\tilde{A}}$ is the exchange of the columns i and j. Observe that $S_{ij}S_{ij} = \mathcal{I}$ results in the identity matrix and that permutations matrices are symmetric, in the sense that $S_{ij}^T = S_{ij}$.

With that said if we consider a series of multiplication on permutation matrix we have that

$$P = S_{ij}S_{kl}\dots S_{mp} \qquad \Rightarrow \qquad P^T = S_{mp}^T\dots S_{kl}^TS_{ij}^T = S_{mp}\dots S_{kl}S_{ij}$$

because we have that $P^TP = \mathcal{I}$. In general a permutation matrix if a series of product of exchanges collected in a single matrix P such that $P^{-1} = P^T$.

Jacobi modification The main idea of the LU decomposition is to find $n-1$ matrices $\mathbf{L}_i$ such that $\mathbf{L}_{n-1}\dots\mathbf{L}_2\mathbf{L}_1\mathbf{PAQ} = \mathbf{U}$ (where $\mathbf{L} = (\mathbf{L}_{n-1}\dots\mathbf{L}_2\mathbf{L}_1)^{-1}$), where $\mathbf{U}$ is upper triangular and L_i are all lower ones. The **Jacobi modification** leverage the same idea, but the permuted matrix $\mathbf{PAQ}$ is pre-multiplied by a series of Jordan normal matrices $\mathbf{J}_i$ resulting in a matrix of the form:

$$\mathbf{J}_n\dots\mathbf{J}_2\mathbf{J}_1\mathbf{PAQ} = \left[\begin{array}{c|c} \mathbf{I} & \\ \hline 0 & \end{array}\right]$$

where the *blank spaces* can be filled with non-zero elements.

6.2 DAE index and index reduction

As already discussed, a linear differential algebraic equation can be regarded as $\mathbf{E}\boldsymbol{y}' + \mathbf{A}\boldsymbol{y} = \boldsymbol{c}(t)$ (where we assume that the pair $(\mathbf{E}, \mathbf{A})$ is a regular pencil), then with the Kronecker decomposition we had the formulation shown in equation 6.9 (page 91). That allowed to re-state the original DAE problem into an algebraic part $\boldsymbol{\alpha} = \sum_{j=0}^{p-1}(-1)^j\mathbf{N}^j\boldsymbol{d}^{(j)}$ and an ordinary differential one $\boldsymbol{\beta}'(t) + \mathbf{J}\boldsymbol{\beta}(t)\boldsymbol{e}(t)$. Deriving the algebraic part evaluates to $\boldsymbol{\alpha}'(t) = \sum_{j=0}^{p-1}(-\mathbf{N})^j\boldsymbol{d}^{(j+1)}(t)$, meaning that the Kronecker normal form allows to transform the original DAE problem into a system of ordinary differential equations:

$$\begin{pmatrix}\boldsymbol{\alpha}'\\ \boldsymbol{\beta}'\end{pmatrix} + \begin{bmatrix}0 & 0\\ 0 & \mathbf{J}\end{bmatrix}\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{pmatrix} = \begin{pmatrix}\boldsymbol{e}(t)\\ \sum_{j=0}^{p-1}(-\mathbf{N})^j\boldsymbol{d}^{(j+1)}(t)\end{pmatrix} \tag{6.12}$$

We can observe so that starting from a linear differential algebraic equation with constant coefficients $\mathbf{E}\boldsymbol{y}' + \mathbf{A}\boldsymbol{y} = \boldsymbol{c}(t)$ we have that after p derivation (where p is the nilpotency order of $\mathbf{N}$ in the Kronecker normal form) and *some algebraic manipulation* we obtain an ordinary differential equation; as definition we say that the DAE has a **Kronecker index** p.

Differential index The minimum number of derivations of the system $\mathbf{E}\boldsymbol{y}' + \mathbf{A}\boldsymbol{y} = \boldsymbol{c}(t)$ required to transform the DAE into an ODE is called **differential index**.

Assuming the general expression $\boldsymbol{F}(\boldsymbol{y}, \boldsymbol{y}', t) = 0$, but also the derivatives $\frac{d}{dt}\boldsymbol{F}(\boldsymbol{y}, \boldsymbol{y}', t) = \frac{\partial \boldsymbol{F}}{\partial \boldsymbol{y}}\boldsymbol{y}' \frac{\partial \boldsymbol{F}}{\partial \boldsymbol{y}'}\boldsymbol{y}'' + \frac{\partial \boldsymbol{F}}{\partial t}$, $\frac{d^2}{dt^2}\boldsymbol{F}$ up to the p-th order $\frac{d^p}{dt^p}\boldsymbol{F}$, such values can be combined to obtain an expression in the form $\boldsymbol{y}' = \boldsymbol{G}(\boldsymbol{y}, t)$. The differential index is the minimum number of derivations p required to transform $\boldsymbol{F}(\boldsymbol{y}, \boldsymbol{y}', t) = 0$ into $\boldsymbol{g}' = \boldsymbol{G}(\boldsymbol{y}, t)$.
As special cases

- if the Kronecker normal form is of type $\mathbf{PEQ} = \mathbf{I}$ and $\mathbf{PAQ} = \mathbf{J}$, then we have that

$$\begin{aligned} \mathbf{PE}\boldsymbol{y}' + \mathbf{PA}\boldsymbol{y} &= \mathbf{P}\boldsymbol{c} \\ \mathbf{PEQQ}^T\boldsymbol{y}' + \mathbf{PAQQ}^T\boldsymbol{y} &= \mathbf{P}\boldsymbol{c} \qquad \leftarrow \quad \boldsymbol{z} = \mathbf{Q}^T\boldsymbol{y} \\ \mathbf{I}\boldsymbol{z}' + \mathbf{J}\boldsymbol{z} &= \mathbf{P}\boldsymbol{c} \\ \Rightarrow \qquad \boldsymbol{z}' &= \mathbf{P}\boldsymbol{c} - \mathbf{J}\boldsymbol{z} \end{aligned}$$

  This is a purely ordinary differential equation.

- alternatively if the Kronecker normal form is such that $\mathbf{PEQ} = \mathbf{N}$ and $\mathbf{PAQ} = \mathbf{I}$, then we have

$$\begin{aligned} \mathbf{PEQQ}^T\boldsymbol{y}' + \mathbf{PAQQ}^T\boldsymbol{y} &= \mathbf{P}\boldsymbol{c} \qquad \leftarrow \quad \boldsymbol{z} = \mathbf{Q}^T\boldsymbol{y} \\ \mathbf{N}\boldsymbol{z}' + \boldsymbol{z} &= \mathbf{P}\boldsymbol{c} = \boldsymbol{d} \end{aligned}$$

  Knowing that $\mathbf{N}$ is a nilpotent matrix of order p, recalling previous *tricks we have that*

$$\boldsymbol{z} = \sum_{j=0}^{p-1}(-\mathbf{N})^j\boldsymbol{d}^{(j)}$$

  This is a purely algebraic equation.

Example 6.3: DAE to ODE

Considering the differential algebraic equation

$$\begin{cases}\dot{x}_1 + \dot{x}_2 + x_1 = 1\\ \dot{x}_1 + \dot{x}_2 + x_1 + x_2 = t\end{cases}$$

subtracting from the second equation the first one results in $x_2 = t - 1$ that derived determines $\dot{x}_2 = 1$: we have so $\dot{x}_2$ expressed as an ordinary differential equation. Substituting this result in the first equation we have $\dot{x}_1 + 1 + x_1 = 0$, hence $\dot{x}_1 = -x_1$. After 1 derivation the resulting ODE is

$$\begin{cases} \dot{x}_1 = -x_1 \\ \dot{x}_2 = 1 \end{cases}$$

The differential index in this case is 1 (because we only derived $\dot{x}_2 = t - 1$ once).

Systematic index reduction algorithm Given a linear DAE with constant coefficients $\mathbf{E}\boldsymbol{y}' + \mathbf{A}\boldsymbol{y} = \boldsymbol{c}$ (where usually $\mathbf{E}$ is singular), we can find two permutation matrices $\mathbf{P}, \mathbf{Q}$ in order to have the factorization $\mathbf{PEQ} = \mathbf{LU}$ where $\mathbf{L}$ is lower triangular and $\mathbf{Q}$ is *upper trapezoidal* due to singularity of E. Knowing that $\mathbf{Q}^{-1} = \mathbf{Q}^T$, we also have that $\mathbf{PE} = \mathbf{LUQ}^T$: this last permutation $\mathbf{Q}^T$ corresponds simply to column commutations, meaning that we can regard

$$\mathbf{PE} = \mathbf{LUQ}^T = \mathbf{LM}$$

where $\mathbf{M}$ is a matrix that in general is composed in two *vertically stacked rectangular blocks*: the upper one that' generally non zero and the lower one identically null, hence $\mathbf{M} = \left[\begin{array}{c} \mathbf{M}_1 \neq 0 \\ \hline \mathbf{M}_2 = 0 \end{array}\right]$. With this consideration we can pre-multiply the linear DAE by $\mathbf{L}^{-1}\mathbf{P}$ what we obtain is the formulation

$$\begin{aligned} \mathbf{L}^{-1}\mathbf{PE}\boldsymbol{y}' + \mathbf{L}^{-1}\mathbf{PA}\boldsymbol{y} &= \mathbf{L}^{-1}\mathbf{P}\boldsymbol{c} \\ \mathbf{M}\boldsymbol{y}' + \mathbf{N}\boldsymbol{y} &= \boldsymbol{d} \end{aligned} \tag{6.13}$$

Splitting the matrices/vector $\mathbf{M}, \mathbf{N}, \boldsymbol{d}$ in order to *match* the dimension of the blocks of $\mathbf{M}$, then we can consider the initial differential algebraic equations as made of

$$\begin{cases} \mathbf{M}_1\boldsymbol{y}' + \mathbf{N}_1 = \boldsymbol{d}_1 & : \text{differential part} \\ \mathbf{N}_2\boldsymbol{y} = \boldsymbol{d}_2 & : \text{algebraic part} \end{cases}$$

The easiest thing to do to reduce the DAE into a system of ordinary differential equation is to derive in time the algebraic part, that evaluates to $\mathbf{N}_2\boldsymbol{y}' = \boldsymbol{d}_2'$: with that we obtain a *pure* system of ordinary differential equations in the form

$$\begin{cases} \mathbf{M}_1\boldsymbol{y}' + \mathbf{N}_1 = \boldsymbol{d}_1 & : \text{differential part} \\ \mathbf{N}_2\boldsymbol{y}' = \boldsymbol{d}_2' & : \text{algebraic part} \end{cases}$$

This results in a *new* differential algebraic system in the form $\mathbf{E}_1\boldsymbol{y}' + \mathbf{A}_1\boldsymbol{y} = \boldsymbol{e}_1$ where

$$\mathbf{E}_1 = \begin{bmatrix} \mathbf{M}_1 \\ \mathbf{N}_2 \end{bmatrix} \qquad \mathbf{A}_1 = \begin{bmatrix} \mathbf{N}_1 \\ 0 \end{bmatrix} \qquad \boldsymbol{e}_1 = \begin{pmatrix} \boldsymbol{d}_1 \\ \boldsymbol{d}_2' \end{pmatrix}$$

The so computed matrix $\mathbf{E}_1$ can still be singular, but we can apply this same algorithm until the obtained system is non-singular, hence solvable: the number of required derivations is the **index** of the differential algebraic equation.

Example 6.4: index computation and reduction

Considering a differential algebraic system in the form

$$\begin{cases} \dot{x}_1 + \dot{x}_2 + \dot{x}_3 + x_1 = \sin t \\ \dot{x}_1 + \dot{x}_2 + \dot{x}_3 + x_3 = t \\ x_1 + x_3 = \cos t \end{cases}$$

in order to perform the index reduction we can rewrite the system in a more *compact* form in order to perform more easily the Gauss-Jordan solution of the system:

$$\begin{array}{cccccccc} \dot{x}_1 & \dot{x}_2 & \dot{x}_3 & x_1 & x_2 & x_3 & \text{RHS} \\ \left[\begin{array}{ccc|ccc|c} 1 & 1 & 1 & 1 & 0 & 0 & \sin t \\ 1 & 1 & 1 & 0 & 0 & 1 & t \\ 0 & 0 & 0 & 1 & 0 & 1 & \cos t \end{array}\right] \end{array}$$

Performing the transformation on the rows of such matrix $(2) \mapsto (2) - (1)$ we obtain the matrix

$$\left[\begin{array}{ccc|ccc|c} 1 & 1 & 1 & 1 & 0 & 0 & \sin t \\ 0 & 0 & 0 & -1 & 0 & 1 & t - \sin t \\ 0 & 0 & 0 & 1 & 0 & 1 & \cos t \end{array}\right]$$

Observing that the last two rows are identically zero for what concerns the entries in $\mathbf{E}$, then we can derive such algebraic part by *shifting* to the left the block $\mathbf{N}_2$ and deriving in time the last column associated to the right hand side $\boldsymbol{d}$, obtaining

$$\left[\begin{array}{ccc|ccc|c} 1 & 1 & 1 & 1 & 0 & 0 & \sin t \\ -1 & 0 & 1 & 0 & 0 & 0 & 1 - \cos t \\ 1 & 0 & 1 & 0 & 0 & 0 & -\sin t \end{array}\right]$$

Continuing the Gauss-Jordan reduction

$$\xrightarrow[(3)\mapsto(3)-(1)]{(2)\mapsto(2)+(1)} \left[\begin{array}{ccc|ccc|c} 1 & 1 & 1 & 1 & 0 & 0 & \sin t \\ 0 & 1 & 2 & 1 & 0 & 0 & 1 - \cos t + \sin t \\ 0 & -1 & 0 & -1 & 0 & 0 & -2\sin t \end{array}\right]$$

$$\xrightarrow[(3)\mapsto(3)+(2)]{(1)\mapsto(1)-(2)} \left[\begin{array}{ccc|ccc|c} 1 & 0 & -1 & 0 & 0 & 0 & \cos t - 1 \\ 0 & 1 & 2 & 1 & 0 & 0 & 1 - \cos t + \sin t \\ 0 & 0 & 2 & 0 & 0 & 0 & 1 - \cos t - \sin t \end{array}\right]$$

$$\xrightarrow[(2)\mapsto(2)-(3)]{(1)\mapsto(1)+\frac{1}{2}(3)} \left[\begin{array}{ccc|ccc|c} 1 & 0 & 0 & 0 & 0 & 0 & \frac{\cos t - \sin t - 1}{2} \\ 0 & 1 & 0 & 1 & 0 & 0 & 2\sin t \\ 0 & 0 & 2 & 0 & 0 & 0 & 1 - \cos t - \sin t \end{array}\right]$$

After all this reductions we obtained a non-singular matrix $\mathbf{E}$ determined just after one differentiation of the original problem (hence the index of the DAE is $p = 1$) and the resulting ordinary differential equation is

$$\begin{cases} \dot{x}_1 = \frac{1}{2}(\cos t - \sin t - 1) \\ \dot{x}_2 = 2\sin t - x_1 \\ \dot{x}_3 = \frac{1}{2}(1 - \cos t - \sin t) \end{cases}$$

When the algorithm fails Considering now a differential algebraic system characterized by the equations

$$\begin{cases} x' + y' + z' + w' + x = t \\ x' - x = t^2 \\ y' - x = t \\ x' + y' = 0 \end{cases} \quad \leftrightarrow \quad \left[\begin{array}{cccc|cccc|c} 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & t \\ 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & t^2 \\ 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & t \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

after some steps in the Jordan-Gauss reduction for linear systems we obtain the form

$$\left[\begin{array}{cccc|cccc|c} 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & t^2 \\ 0 & -1 & -1 & -1 & -2 & 0 & 0 & 0 & t^2 - t \\ 0 & 0 & -1 & -1 & -1 & 0 & 0 & 0 & -t \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & t - t^2 \end{array}\right]$$

Clearly the last row is a contradiction, in fact what it says is that $0 = t - t^2$ that's not verified in general. This is due to the fact that the initial choice of the pair $(\mathbf{A}, \mathbf{E})$ was **not** a **regular pencil**: we in fact have that

$$\det(\mathbf{E} - \lambda\mathbf{A}) = \begin{vmatrix} 1-\lambda & 1 & 1 & 1 \\ 1+\lambda & 0 & 0 & 0 \\ 1+\lambda & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{vmatrix} = - \begin{vmatrix} 1+\lambda & 0 & 0 \\ 1+\lambda & 0 & 0 \\ 1 & 1 & 0 \end{vmatrix} = 0$$

where the expansions used to compute the determinant are made along the last column on each submatrix.

Linear DAE with non-constant coefficients Considering the more general case of non-constant coefficient for linear differential algebraic equation, hence in the form $\mathbf{E}(t)\boldsymbol{y}' + \mathbf{A}\boldsymbol{y} = \boldsymbol{c}$, we might want to know if the condition on the regular pencil $(\mathbf{A}, \mathbf{E})$ must still be required. Considering the simple system $\mathbf{E}(t) = \begin{bmatrix} 1 & t \\ 0 & 0 \end{bmatrix}$ and $\mathbf{A}(t) = \begin{bmatrix} 0 & 0 \\ 1 & t \end{bmatrix}$ we can show that they are not regular pencil, in fact

$$\det(\mathbf{E} - \lambda\mathbf{A}) = \begin{vmatrix} 1 & t \\ 1-\lambda & t-\lambda \end{vmatrix} = t - \lambda - t(1-\lambda) = 0$$

However writing explicitly the differential algebraic equation we have

$$\begin{cases} x' + ty' = c_1 & : \text{differential part} \\ x + ty = c_2 & : \text{algebraic part} \end{cases}$$

Deriving in time the algebraic equation results in the system

$$\begin{cases} x' + ty' = c_1 \\ x' + ty' + y = c_2' \end{cases} \xrightarrow{(2)\mapsto(2)-(1)} \begin{cases} x' + ty' = c_1 \\ y = c_2' - c_1 \end{cases}$$

Deriving one more time the second equation (that's algebraic) allows to compute a *non-singular* system of ODEs characterized by solutions

$$y' = c_2'' - c_1' \qquad x' = c_1 - (c_2'' - c_1')t$$

The index of the original DAE is so 2.
In general not being regular pencil for non-constant linear DAEs is not a problem; in fact the the system $\mathbf{E}(t)\boldsymbol{x}' + \mathbf{A}(t)\boldsymbol{x} = \boldsymbol{c}$, after some algebraic manipulation, can be reduced to a form

$$\begin{bmatrix} \tilde{\mathbf{E}}_1(t) \\ 0 \end{bmatrix} \boldsymbol{x}' + \begin{bmatrix} \tilde{\mathbf{A}}_1(t) \\ \tilde{\mathbf{A}}_2(t) \end{bmatrix} \boldsymbol{x} = \begin{pmatrix} \tilde{c}_1 \\ \tilde{c}_2 \end{pmatrix}$$

Deriving the algebraic part this time doesn't mean just *shifting* $\tilde{\mathbf{A}}_2(t)$ on the left to complex the matrix $\tilde{\mathbf{E}}$, but also introduces the derivative of the entries of $\tilde{\mathbf{A}}_2(t)$:

$$\begin{bmatrix} \tilde{\mathbf{E}}_1(t) \\ \tilde{\mathbf{A}}_2(t) \end{bmatrix} \boldsymbol{x}' + \begin{bmatrix} \tilde{\mathbf{A}}_1(t) \\ \tilde{\mathbf{A}}_2'(t) \end{bmatrix} \boldsymbol{x} = \begin{pmatrix} \tilde{c}_1 \\ \tilde{c}_2' \end{pmatrix}$$

This makes the system in general non-singular, having the possibility always to have a solution.

Tricks in the computation Recalling example 6.4, after writing the linear system we described it in a *unique* matrix containing both matrices $\mathbf{E}, \mathbf{A}$ and the right-hand side of the equation. This can be done for linear system, however we can note that the Gauss reduction is performed on the matrix $\mathbf{E}$ in order to obtain a form $\begin{bmatrix} \mathbf{I} & \\ & \mathbf{0} \end{bmatrix}$. The general idea is so to compute the Gauss steps on the system

$$\left[\begin{array}{c|c} \mathbf{E} & \mathbf{I} \end{array}\right] \xrightarrow{\text{Gauss reduction}} \left[\begin{array}{c|c} \mathbf{T}_m \dots \mathbf{T}_2\mathbf{T}_1\mathbf{E} & \mathbf{T}_m \dots \mathbf{T}_2\mathbf{T}_1\mathbf{I} \end{array}\right] = \left[\begin{array}{cc|c} \mathbf{I} & & \\ & \mathbf{0} & \mathbf{T} \end{array}\right]$$

With such transformation matrix $\mathbf{T}$ computing $\mathbf{TE}\boldsymbol{y}' + \mathbf{TA}\boldsymbol{y} = \mathbf{T}\boldsymbol{c}$ results in a separation of the differential part from the algebraic one and we can apply the algorithm (by deriving the algebraic part and iterate).

Example 6.5: simple pendulum and index reduction

Recalling the simple pendulum described at the start of this chapter, we retrieved the differential algebraic equation describing the system as

$$\begin{cases} x' = u \\ y' = v \\ my' + \lambda x = 0 \\ mv' + \lambda y = -mg \\ x^2 + y^2 - l^2 = 0 \end{cases}$$

Describing the time-dependent coordinates in the vector $\boldsymbol{z}' = (x, y, u, v, \lambda)$, the first thing to to is to re-state the problem in a form $\mathbf{E}(\boldsymbol{z}, t)\boldsymbol{z}' = \boldsymbol{G}(\boldsymbol{z}, t)$, so

$$\begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & m & & \\ & & & m & \\ & & & & 0 \end{bmatrix} \begin{pmatrix} x' \\ y' \\ u' \\ v' \\ \lambda' \end{pmatrix} = \begin{pmatrix} u \\ v \\ -\lambda x \\ -\lambda y - mg \\ l^2 - x^2 - y^2 \end{pmatrix}$$

Exploiting the yet-described trick we consider the linear system

$$\left[\begin{array}{ccccc|ccccc} 1 & & & & & 1 & & & & \\ & 1 & & & & & 1 & & & \\ & & m & & & & & 1 & & \\ & & & m & & & & & 1 & \\ & & & & 0 & & & & & 1 \end{array}\right]$$

The first step of the index reduction is quite simple and consists in the multiplication by $\frac{1}{m}$ of both the 3^{rd} and 4^{th} rows, resulting in

$$\left[\begin{array}{cc|c} \mathbf{I} & & \\ & \mathbf{0} & \mathbf{T}_1 \end{array}\right] = \left[\begin{array}{ccccc|ccccc} 1 & & & & & 1 & & & & \\ & 1 & & & & & 1 & & & \\ & & 1 & & & & & \frac{1}{m} & & \\ & & & 1 & & & & & \frac{1}{m} & \\ & & & & 0 & & & & & 1 \end{array}\right]$$

Applying the transformation matrix $\mathbf{T}$ on the initial system determines so the form

$$\begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & 1 & \\ & & & & 0 \end{bmatrix} \begin{pmatrix} x' \\ y' \\ u' \\ v' \\ \lambda' \end{pmatrix} = \mathbf{T}_1\boldsymbol{G} = \begin{pmatrix} u \\ v \\ -\frac{\lambda}{m}x \\ -\frac{\lambda}{m}y - g \\ l^2 - x^2 - y^2 \end{pmatrix} = \boldsymbol{G}_1$$

Deriving with respect to time the algebraic equation (that's only the last row) evaluates to $\frac{d}{dt}(l^2 - x^2 - y^2) = -2xx' - 2yy'$; this determines the new system

$$\begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & 1 & \\ -2x & -2y & 0 & 0 & 0 \end{bmatrix} \begin{pmatrix} x' \\ y' \\ u' \\ v' \\ \lambda' \end{pmatrix} = \begin{pmatrix} u \\ v \\ -\frac{\lambda}{m}x \\ -\frac{\lambda}{m}y - g \\ 0 \end{pmatrix}$$

We can apply the same method to reduce this system, by so determining a second transformation matrix $\mathbf{T}_2$ for this system using the Gauss reduction:

$$\left[\begin{array}{ccccc|ccccc} 1 & & & & & 1 & & & & \\ & 1 & & & & & 1 & & & \\ & & 1 & & & & & 1 & & \\ & & & 1 & & & & & 1 & \\ -2x & -2y & 0 & 0 & 0 & & & & & 1 \end{array}\right]$$

$$\xrightarrow{(5)\mapsto(5)+2x(1)+2y(1)} \left[\begin{array}{ccccc|ccccc} 1 & & & & & 1 & & & & \\ & 1 & & & & & 1 & & & \\ & & 1 & & & & & 1 & & \\ & & & 1 & & & & & 1 & \\ 0 & 0 & 0 & 0 & 0 & 2x & 2y & 0 & 0 & 1 \end{array}\right]$$

This new transformation matrix $\mathbf{T}_2$ determines a new system

$$\begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & 1 & \\ & & & & 0 \end{bmatrix} \begin{pmatrix} x' \\ y' \\ u' \\ v' \\ \lambda' \end{pmatrix} = \mathbf{T}_2\boldsymbol{G}_1 = \begin{pmatrix} u \\ v \\ -\frac{\lambda}{m}x \\ -\frac{\lambda}{m}y - g \\ 2xu + 2yv \end{pmatrix} = \boldsymbol{G}_2$$

Deriving the algebraic part $\frac{d}{dt}(2xu + 2yv) = 2x'u + 2xu' + 2y'v + 2yv'$ determines a new differential matrix $\mathbf{E}_3$ that can so be reduced using the Gauss method:

$$\left[\begin{array}{ccccc|ccccc} 1 & & & & & 1 & & & & \\ & 1 & & & & & 1 & & & \\ & & 1 & & & & & 1 & & \\ & & & 1 & & & & & 1 & \\ 2u & 2v & 2x & 2y & 0 & & & & & 1 \end{array}\right]$$

$$\xrightarrow{(5)\mapsto(5)-2u(1)-2v(1)-2x(3)-2y(4)} \left[\begin{array}{ccccc|ccccc} 1 & & & & & 1 & & & & \\ & 1 & & & & & 1 & & & \\ & & 1 & & & & & 1 & & \\ & & & 1 & & & & & 1 & \\ 0 & 0 & 0 & 0 & 0 & -2u & -2v & -2x & -2y & 1 \end{array}\right]$$

determining

$$\begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & 1 & \\ & & & & 0 \end{bmatrix} \begin{pmatrix} x' \\ y' \\ u' \\ v' \\ \lambda' \end{pmatrix} = \mathbf{T}_3\boldsymbol{G}_2 = \begin{pmatrix} u \\ v \\ -\frac{\lambda}{m}x \\ -\frac{\lambda}{m}y - g \\ -2u^2 - 2v^2 + 2\frac{\lambda}{m}x^2 + 2\frac{\lambda}{m}y^2 + 2yg \end{pmatrix} = \boldsymbol{G}_3$$

Deriving one more time the lonely algebraic equation results in

$$\frac{d}{dt}\left(-2u^2-2v^2+2\frac{\lambda}{m}x^2+2\frac{\lambda}{m}y^2+2yg\right)=2\lambda'\frac{x^2+y^2}{m}+4\lambda\frac{xx'+yy'}{m}-4uu'-4vv'+2y'g$$

We so reduce the system

$$\left[\begin{array}{ccccc|ccccc} 1 & & & & & 1 & & & & \\ & 1 & & & & & 1 & & & \\ & & 1 & & & & & 1 & & \\ & & & 1 & & & & & 1 & \\ & & & & 1 & & & & & 1 \\ 4\frac{\lambda}{m}x & 4\frac{\lambda}{m}y+2g & -4u & -4v & 2\frac{x^2+y^2}{m} & & & & & \end{array}\right]$$

$$\xrightarrow{(5)\mapsto(5)-4\frac{\lambda}{m}x(1)-\left(4\frac{\lambda}{m}y+2g\right)(2)+4u(3)+4v(4)} \left[\begin{array}{ccccc|ccccc} 1 & & & & & 1 & & & & \\ & 1 & & & & & 1 & & & \\ & & 1 & & & & & 1 & & \\ & & & 1 & & & & & 1 & \\ 0 & 0 & 0 & 0 & 2\frac{x^2+y^2}{m} & -4\frac{\lambda}{m}x & -4\frac{\lambda}{m}y-2g & 4u & 4v & 1 \end{array}\right]$$

$$\xrightarrow{(5)\mapsto\frac{m}{2(x^2+y^2)}(5)} \left[\begin{array}{ccccc|ccccc} 1 & & & & & 1 & & & & \\ & 1 & & & & & 1 & & & \\ & & 1 & & & & & 1 & & \\ & & & 1 & & & & & 1 & \\ 0 & 0 & 0 & 0 & 1 & -\frac{2\lambda x}{x^2+y^2} & -\frac{2\lambda y}{x^2+y^2}-\frac{mg}{x^2+y^2} & \frac{2mu}{x^2+y^2} & \frac{2mv}{x^2+y^2} & \frac{m}{2(x^2+y^2)} \end{array}\right]$$

Finally we have a transform matrix **E** that's non-singular and the ordinary differential equation originated from the initial DAE is so

$$\begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & 1 & \\ & & & & 1 \end{bmatrix}\begin{pmatrix} x' \\ y' \\ u' \\ v' \\ \lambda' \end{pmatrix}=\mathbf{T}_4\boldsymbol{G}_3=\begin{pmatrix} u \\ v \\ -\frac{\lambda}{m}x \\ -\frac{\lambda}{m}y-g \\ \frac{-4\lambda(xy+yv)-3mgv}{x^2+y^2} \end{pmatrix}$$

6.2.1 Introduction of dummy variables

Considering the following differential algebraic equation retrieved by a fairly simple mechanical system that's in the form

$$\begin{cases} x_1'=u_1 \\ y_1'=v_1 \\ x_2'=u_2 \\ u_1'=2\lambda_1(x_1-x_2)+2\lambda_2x_1 \\ v_1'=2y_1(\lambda_1-\lambda_2)-y \\ u_2'=2\lambda_2(x_2-x_1) \\ x_1^2+y_1^2-1=0 \\ (x_1-x_2)^2+y_1^2-1=0 \end{cases} \tag{6.14}$$

The difficult part for manually solving this problems lies in the implicit differentiation of the right hand sides of the ODEs: the three ones are quite simple (in fact $\frac{d}{dt}x_1'=u_1'$) while the others are more complex ($\frac{d}{dt}u_2'=2\lambda_2'(x_2-x_1)+2\lambda_2(x_2'-x_1')$) and so is not in general a good idea to perform the index reduction on such system (because each steps requires the differentiation in time, increasing the complexity of the calculations).

The idea is so to introduce some ***dummy variables*** in the form $\dot{z}$ that allows to rewrite the second three ODEs as

$$u_1' = \dot{u}_1 \qquad v_1' = \dot{v}_1 \qquad u_2' = \dot{u}_2$$

subjected to the following algebraic constraints:

$$0 = 2\lambda_1(x_1 - x_2) + 2\lambda_2 x_1 - \dot{u}_1 \qquad 0 = 2y_1(\lambda_1 - \lambda_2) - y - \dot{v}_1 \qquad 0 = 2\lambda_2(x_2 - x_1) - \dot{u}_2$$

With this idea the initial differential algebraic equation can be regarded as

$$\begin{cases} \left.\begin{aligned} x_1' &= u_1 \\ y_1' &= v_1 \\ x_2' &= u_2 \\ u_1' &= \dot{u}_1 \\ v_1' &= \dot{v}_1 \\ u_2' &= \dot{u}_2 \end{aligned}\right\} & \text{: simpler ODE part} \\ \left.\begin{aligned} 0 &= 2\lambda_1(x_1 - x_2) + 2\lambda_2 x_1 - \dot{u}_1 \\ 0 &= 2y_1(\lambda_1 - \lambda_2) - y - \dot{v}_1 \\ 0 &= 2\lambda_2(x_2 - x_1) - \dot{u}_2 \end{aligned}\right\} & \text{: additional algebraic constraint} \\ \left.\begin{aligned} 0 &= x_1^2 + y_1^2 - 1 \\ 0 &= (x_1 - x_2)^2 + y_1^2 - 1 \end{aligned}\right\} & \text{: original algebraic constraint} \end{cases} \tag{6.15}$$

In order to reduce the index we can so differentiate in time the algebraic constraints: starting off with the newly added one what we obtain is (mathematical simplification are already performed)

$$\begin{aligned} \frac{d}{dt}\left(2\lambda_1(x_1 - x_2) + 2\lambda_2 x_1 - \dot{u}_1\right) &= 2\lambda_1'(x_1 - x_2) + 2\lambda_1(u_1 - u_2) + 2\lambda_2' x_1 + 2\lambda_2 u_1 + \dot{u}_1' \\ \frac{d}{dt}\left(2y_1(\lambda_1 - \lambda_2) - y - \dot{v}_1\right) &= 2v_1(\lambda_1 - \lambda_2) + 2y_1(\lambda_1' - \lambda_2') - \dot{v}_1' \\ \frac{d}{dt}\left(2\lambda_2(x_2 - x_1) - \dot{u}_2\right) &= 2\lambda_2'(x_2 - x_1) + 2\lambda_2(u_2 - u_1) - \dot{u}_2' \end{aligned}$$

In this case the differentiation already presents differential terms (in the variables $\dot{u}_1', \dot{v}_1', \dot{u}_2', \lambda_1', \lambda_2'$) and so *it doesn't make sense* to continue with the differentiation in time of this expressions. Differentiating instead the *original* algebraic constraints what we obtain is

$$\frac{d}{dt}\left(x_1^2 + y_1^2\right) = 2x_1 x_1' + 2y_1 y_1' = 2x_1 u_1 + 2y_1 v_1 \tag{a}$$

$$\frac{d}{dt}\left((x_1 - x_2)^2 + y_1^2 - 1\right) = 2(x_1 - x_2)(x_1' - x_2') + 2y_1 y_1' = 2(x_1 - x_2)(u_1 - u_2) + 2y_1 v_1 \tag{b}$$

This equations (after substituting all the variables $x_i' = u_i$) presents no differential part, hence we have to reduce one more time the index by differentiating only this two algebraic equations in time:

$$\frac{d}{dt}(a) = 2u_1^2 + 2x_1\dot{u}_1 + 2v_1^2 + 2y_1\dot{v}_1 \tag{c}$$

$$\frac{d}{dt}(b) = 2(u_1 - u_2)^2 + 2(x_1 - x_2)(\dot{u}_1 - \dot{u}_2) + 2v_1^2 + 2y_1\dot{v}_1 \tag{d}$$

We still need to reduce the index (because this expression are still algebraic) and after this more differentiation we finally obtain an ordinary differential equation:

$$\frac{d}{dt}(c) = 6u_1\dot{u}_1 + 6v_1\dot{v}_1 + 2x_1\dot{u}_1' + 2y_1\dot{v}_1'$$
$$\frac{d}{dt}(d) = 6(u_1-u_2)(\dot{u}_1-\dot{u}_2) + 6v_1\dot{v}_1 + 2(x_1-x_2)(\dot{u}_1'-\dot{u}_2') + 2y_1\dot{v}_1'$$

Ended the index reduction used to transform the algebraic constraints in ordinary differential part, we can rewrite such obtained ODEs in a matrix form as

$$\underbrace{\begin{bmatrix} 2(x_1-x_2) & 2x_1 & -1 & 0 & 0 \\ 2y_1 & -2y_1 & 0 & -1 & 0 \\ 0 & -2(x_1-x_2) & 0 & 0 & -1 \\ 0 & 0 & 2x_1 & 2y_1 & 0 \\ 0 & 0 & 2(x_1x_2) & 2y_1 & -2(x_1-x_2) \end{bmatrix}}_{\mathbf{E}} \underbrace{\begin{pmatrix} \lambda_1' \\ \lambda_2' \\ \dot{u}_1' \\ \dot{v}_1' \\ \dot{u}_2' \end{pmatrix}}_{\boldsymbol{z}'} = \underbrace{\begin{pmatrix} 2\lambda_2(u_2-u_1) - 2\lambda_2 u_1 \\ 2v_1(\lambda_2-\lambda_1) \\ 2\lambda_2(u_1-u_2) \\ -6u_1\dot{u}_1 - 6v_1\dot{v}_1 \\ -6(u_1-u_2)(\dot{u}_1-\dot{u}_2) - 6v_1\dot{v}_1 \end{pmatrix}}_{\boldsymbol{G}}$$

After all this steps we can so consider that the equivalent ordinary differential system of the initial DAE problem reported in equation 6.14 is the one determined as

$$\begin{cases} x_1' = u_1 \\ y_1' = v_1 \\ x_2' = u_2 \\ u_1' = \dot{u}_1 \\ v_1' = \dot{v}_1 \\ u_2' = \dot{u}_2 \\ \mathbf{E}\boldsymbol{z}' = \boldsymbol{G} \end{cases} \tag{6.16}$$

The initial condition of such ordinary differential equation satisfy the hidden *original* constraints

$$0 = x_1^2 + y_1^2 - 1 \qquad 0 = (x_1-x_2)^2 + y_1^2 - 1$$
$$0 = 2\lambda_1(x_1-x_2) + 2\lambda_2 x_1 - \dot{u}_1 \qquad 0 = 2y_1(\lambda_1-\lambda_2) - y - \dot{v}_1 \qquad 0 = 2\lambda_2(x_2-x_1) - \dot{u}_2$$

but also the expression retrieved while performing the index reduction on the system $(a), (b), (c), (d)$, so

$$2x_1u_1 + 2y_1v_1 = 0 \qquad 2(x_1-x_2)(u_1-u_2) + 2y_1v_1 = 0$$
$$2u_1^2 + 2x_1\dot{u}_1 + 2v_1^2 + 2y_1\dot{v}_1 = 0 \qquad 2(u_1-u_2)^2 + 2(x_1-x_2)(\dot{u}_1-\dot{u}_2) + 2v_1^2 + 2y_1\dot{v}_1 = 0$$

6.2.2 Kernel computation and index reduction

Given a differential algebraic equation $\mathbf{E}(t,\boldsymbol{y})\boldsymbol{y}' = \boldsymbol{G}(t,\boldsymbol{y})$, if the matrix $\mathbf{E}$ is singular then it means that exists at least one vector $\boldsymbol{v} \neq 0$ such that $\mathbf{E}\boldsymbol{v} = 0$. This operation is indeed the **kernel computation** of the matrix $\mathbf{E}$. We define the **kernel** of such matrix the sub-space described by the set

$$\ker\{\mathbf{E}\} = \{\boldsymbol{v} \text{ such that } \mathbf{E}\boldsymbol{v} = 0\} = \text{span}\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_k\}$$

where the vectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$ are determining a basis of $\ker\{\mathbf{E}\}$. Computing the kernel of $\mathbf{E}^T$ evaluates to $\ker\{\mathbf{E}^T\} = \{\boldsymbol{w} \mid \boldsymbol{w}^T\mathbf{E} = 0\} = \text{span}\{\boldsymbol{w}_1^T, \dots, \boldsymbol{w}_p^T\}$. Building so the matrix $\mathbf{K} = \begin{bmatrix} \boldsymbol{w}_1 & \dots & \boldsymbol{w}_p \end{bmatrix}$ (considering all the vectors of the basis of $\ker\{\mathbf{E}^T\}$) as the *concatenations of the vectors* $\boldsymbol{w}_i$ implies that

$$\mathbf{K}^T\mathbf{E} = \begin{bmatrix} \boldsymbol{w}_1^T\mathbf{E} \\ \vdots \\ \boldsymbol{w}_p^T\mathbf{E} \end{bmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

The main idea is that if we are indeed able to compute such matrix $\mathbf{K}$ associated to the kernel of $\mathbf{E}^T$, then we have a way to easily compute the invariant of the differential algebraic equation, in fact

$$\mathbf{K}^T(t,\boldsymbol{y})\Big(\mathbf{E}(t,\boldsymbol{y})\boldsymbol{y}' = \boldsymbol{G}(t,\boldsymbol{y})\Big) \qquad \Rightarrow \qquad \mathbf{0} = \mathbf{K}^T(t,\boldsymbol{y})\boldsymbol{G}(t,\boldsymbol{y})$$

In general $\mathbf{K} \in \mathbb{R}^{n\times p}$ is a rectangular matrix (where $p < n$) and so if we can determine a matrix $\mathbf{L} \in \mathbb{R}^{n\times(p-n)}$ such that $\mathbf{M} = \begin{bmatrix}\mathbf{K} & \mathbf{L}\end{bmatrix} \in \mathbb{R}^{n\times n}$ is non-singular, then we can observe that we can separate the algebraic part of the DAE from the differential part, in fact

$$\begin{aligned}&\mathbf{M}^T\Big(\mathbf{E}\boldsymbol{y}' = \boldsymbol{G}\Big)\\ &\mathbf{M}^T\mathbf{E}\boldsymbol{y}' = \mathbf{M}^T\boldsymbol{G} \qquad \Rightarrow \begin{cases}\cancel{\mathbf{K}^T\mathbf{E}\boldsymbol{y}'} = \mathbf{K}^T\boldsymbol{G}\\ \mathbf{L}^T\mathbf{E}\boldsymbol{y}' = \mathbf{L}^T\boldsymbol{G}\end{cases}\end{aligned}$$

that can be reduced to a form

$$\begin{cases}\tilde{\mathbf{E}}_1\boldsymbol{y}' = \tilde{\boldsymbol{G}}_1 & \text{differential part}\\ 0 = \tilde{\boldsymbol{G}}_2 & \text{algebraic part}\end{cases} \tag{6.17}$$

We have so separated the algebraic part from the differential one, so we can differentiate this second one in order to reduce the index and obtain a system in the form

$$\begin{cases}\tilde{\mathbf{E}}_1\boldsymbol{y}' = \tilde{\boldsymbol{G}}_1\\ \tilde{\mathbf{E}}_1\boldsymbol{y}' = \tilde{\boldsymbol{G}}_1 - \frac{\partial\tilde{\boldsymbol{G}}_2}{\partial\boldsymbol{y}}\boldsymbol{y}' = \frac{\partial\tilde{\boldsymbol{G}}_2}{\partial t}\end{cases}$$

This newly constructed differential algebraic problem can be reduce following the same procedure until when the kernel of $\mathbf{E}^T$ has a null dimension (it is composed only by the zero vector).

Computation of the kernel With this premise being said, we have to find a way to compute the kernel of the square matrix $\mathbf{E}$; this can be achieved, as example, using the LU decomposition. Considering in fact the decomposition $\mathbf{PEQ} = \mathbf{LU}$ (where $\mathbf{P}, \mathbf{Q}$ are permutation matrices), as was previously discussed, can be inverted to a form $\mathbf{E} = \mathbf{P}^{-1}\mathbf{LUQ}^{-1} = \mathbf{P}^{-1}\mathbf{L}M$ where $\mathbf{M}$ can has the *upper rectangle* that's non-null and the lower one that's null (so is in the form $\mathbf{M} = \begin{bmatrix}\mathbf{M}_1\\ 0\end{bmatrix}$); knowing that $\mathbf{L}$ is non-singular (for the definition of the LU decomposition), then also $\mathbf{P}^{-1}\mathbf{L}$ is nonsingular and so we can build the matrix

$$\mathbf{M} = \mathbf{L}^{-1}\mathbf{PE} = \begin{bmatrix}\mathbf{M}_1\\ 0\end{bmatrix}$$

Multiplying by appropriately matrices of the form $\begin{bmatrix}\mathcal{I} & 0\end{bmatrix}$ and $\begin{bmatrix}0 & \mathcal{I}\end{bmatrix}$ this expressions we can observe that

$$\begin{aligned}\underbrace{\begin{bmatrix}\mathcal{I} & 0\end{bmatrix}\mathbf{L}^{-1}\mathbf{P}}_{=\mathbf{R}^T}\mathbf{E} &= \begin{bmatrix}\mathcal{I} & 0\end{bmatrix}\begin{bmatrix}\mathbf{M}_1\\ 0\end{bmatrix} = \mathbf{M}_1 \neq 0\\ \underbrace{\begin{bmatrix}0 & \mathcal{I}\end{bmatrix}\mathbf{L}^{-1}\mathbf{P}}_{=\mathbf{K}^T}\mathbf{E} &= \begin{bmatrix}0 & \mathcal{I}\end{bmatrix}\begin{bmatrix}\mathbf{M}_1\\ 0\end{bmatrix} = 0\end{aligned}$$

This method allowed us to compute the required matrices as $\mathbf{K} = \mathbf{P}^T\mathbf{L}^{-T}\begin{bmatrix}0\\ \mathcal{I}\end{bmatrix}$ and $\mathbf{R} = \mathbf{P}^T\mathbf{L}^{-T}\begin{bmatrix}\mathbf{I}\\ 0\end{bmatrix}$ that *stacked* determined the non-singular matrix $\begin{bmatrix}\mathbf{K} & \mathbf{R}\end{bmatrix} = \mathbf{P}^T\mathbf{L}^{-T}\begin{bmatrix}0 & \mathcal{I}\\ \mathcal{I} & 0\end{bmatrix}$.

6.3 Semi-explicit form

Given a differential algebraic equation in the form $\boldsymbol{F}(\boldsymbol{z}, \boldsymbol{z}', t) = 0$ that can be expressed as

$$\begin{cases} \boldsymbol{x}' &= \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}, t) \\ 0 &= \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y}, t) \end{cases}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are respectively the differential and algebraic variables of the problem and are such that $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y})$. Let us so consider $x \subseteq z$ the variables in $\boldsymbol{F}(\boldsymbol{z}, \boldsymbol{z}', t)$ that are appearing as derivative, then we can defined $\boldsymbol{x}'$ as $\dot{\boldsymbol{x}}$ thus

$$\boldsymbol{F}(\boldsymbol{z}, \boldsymbol{z}', t) = \boldsymbol{F}(\boldsymbol{z}, \dot{\boldsymbol{x}}, t) = \boldsymbol{F}(\boldsymbol{x}, \boldsymbol{y}, \dot{\boldsymbol{x}}, t) = 0$$

This can be rewritten as

$$\begin{cases} \boldsymbol{x}' &= \dot{\boldsymbol{x}} \\ 0 &= \boldsymbol{F}(\boldsymbol{x}, \boldsymbol{y}, \dot{\boldsymbol{x}}, t) \end{cases}$$

Moreover calling $\boldsymbol{w} = (\dot{\boldsymbol{x}}, \boldsymbol{y})$ what we obtain is a **differential algebraic equation** in the so called **semi-explicit form**:

$$\begin{cases} \boldsymbol{x} &= \dot{\boldsymbol{x}} = \boldsymbol{H}(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{t}) \\ 0 &= \boldsymbol{G}(\boldsymbol{x}, \boldsymbol{w}, t) \end{cases} \tag{6.18}$$

Semi-explicit DAE of index 1 Considering a DAE in the form

$$\begin{cases} \boldsymbol{x}' &= \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}, t) \\ 0 &= \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y}, t) \end{cases}$$

then if it is possible to solve $\boldsymbol{g}$ to obtain $\boldsymbol{y}$ so in the form $\boldsymbol{y} = \boldsymbol{H}(\boldsymbol{x}, t)$, then what we have is that $\boldsymbol{g}\big(\boldsymbol{x}, \boldsymbol{H}(\boldsymbol{x}, t), t\big) = 0$ for all values $\boldsymbol{x}, t$, and so we can consider the DAE as

$$\begin{cases} \boldsymbol{x}' &= \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}, t) \\ \boldsymbol{y} &= \boldsymbol{H}(\boldsymbol{x}, t) \end{cases}$$

Differentiating the second equation (the algebraic part) in time determines

$$\boldsymbol{y}' = \frac{d\boldsymbol{H}(\boldsymbol{x}, t)}{dt} = \frac{\partial \boldsymbol{H}(\boldsymbol{x}, t)}{\partial \boldsymbol{x}} \boldsymbol{x}' + \frac{\partial \boldsymbol{H}(\boldsymbol{x}, t)}{\partial t} = \frac{\partial \boldsymbol{H}(\boldsymbol{x}, t)}{\partial \boldsymbol{x}} \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}, t) + \frac{\partial \boldsymbol{H}(\boldsymbol{x}, t)}{\partial t} = \boldsymbol{L}(\boldsymbol{x}, \boldsymbol{y}, t)$$

and so the differential algebraic equation after one differentiation (so with a index reduction) becomes

$$\begin{cases} \boldsymbol{x}' &= \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}, t) \\ \boldsymbol{y}' &= \boldsymbol{L}(\boldsymbol{x}, \boldsymbol{y}, t) \end{cases}$$

In general an ODE in semi-explicit form of index 1 can be solved without performing the derivation by simply considering

$$\boldsymbol{x}' = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}, t) = \boldsymbol{f}\big(\boldsymbol{x}, \boldsymbol{H}(\boldsymbol{x}, t), t\big) = \tilde{\boldsymbol{f}}(\boldsymbol{x}, t) \tag{6.19}$$

6.3.1 Implicit function theorem

Equation 6.19 relies on the fact that from $\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y}, t) = 0$ allows us to extract the map $\boldsymbol{H}(\boldsymbol{x}, t)$ that allows to explicitly compute $\boldsymbol{y}$ (so satisfying $\boldsymbol{g}\big(\boldsymbol{x}, \boldsymbol{H}(\boldsymbol{x}, t), t\big) = 0$), however up to now we cannot be sure that such map $\boldsymbol{H}$ really exists.

Simplified version Considering a function $f(x,y) \in \mathcal{C}^1(\mathbb{R}^2)$ and a point (x_0, y_0) such that $f(x_0, y_0) = 0$ and $\frac{\partial f}{\partial y}(x_0, y_0) \neq 0$, then there exists an open interval $(x_0 - \delta, x_0 + \delta)$ such that

$$y = H(x) \qquad \text{and} \qquad f\big(x, H(x)\big) = 0 \quad \forall x \in (x_0 - \delta, x_0 + \delta)$$

Moreover we have that

$$\frac{dH(x)}{dx} = -\left(\frac{\partial f(x,y)}{\partial y}\right)^{-1} \frac{\partial f(x,y)}{\partial x} \tag{6.20}$$

Considering the simple example of the function $f(x,y) = y^2 - x$, the point $(x_0, y_0) = (1,1)$ satisfies the conditions of the implicit function theorem, in fact

$$f(x_0, y_0) = 0 \qquad \text{and} \qquad \left.\frac{\partial f}{\partial y}\right|_{x=x_0, y=y_0} = \left. 2y \right|_{x=1,y=1} = 2 \neq 0$$

then we are sure that there exists a matrix H that allows to express y as function of x, in fact rewriting $y^2 = x$ and so $y = \pm\sqrt{x}$. The choice of the sign strictly depends on the value of the point (x_0, y_0): in this case we have that $y = H(x) = \sqrt{x}$. In contrary, if we would have chosen the initial point $(x_0, y_0) = (1, -1)$ the condition for the theorem would have been still verified but the resulting map would have been $y = -\sqrt{x}$.
Choosing instead the point $(x_0, y_0) = (0,0)$ would have resulted in the violation of the condition $\frac{\partial f}{\partial y} \neq 0$, thus the theorem cannot be applied: in such point we can't in fact discriminate the sign of the map $y = \pm\sqrt{x}$. However by inverting x and y in the definition, we can define a map $H(y)$ that determines x, in fact

$$f(0,0) = 0 \qquad \text{and} \qquad \left.\frac{\partial f}{\partial x}\right|_{x=x_0, y=y_0} = 1 \neq 0$$

and so $x = H(y) = y^2$.

Implicit function theorem Given a function $\boldsymbol{f} : \mathcal{A} \subseteq \mathbb{R}^{n+m} \to \mathbb{R}^m$ (where n, m are respectively the number of *independent* and *dependent* variables) and a point $(\boldsymbol{x}_0, \boldsymbol{y}_0) \in \mathcal{A}$ (with $\boldsymbol{x}_0 \in \mathbb{R}^n, \boldsymbol{y}_0 \in \mathbb{R}^m$) such that $\boldsymbol{f}(\boldsymbol{x}_0, \boldsymbol{y}_0) = 0$ and the matrix $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{y}}(\boldsymbol{x}_0, \boldsymbol{y}_0)$ is non-singular, then there exists two open sets $\mathcal{U} \subseteq \mathbb{R}^n, \mathcal{V} \subseteq \mathbb{R}^m$ on top of which is defined a map $\boldsymbol{\phi} : \mathcal{U} \to \mathcal{V}$, $\boldsymbol{\phi} \in \mathcal{C}^1(\mathcal{U}, \mathcal{V})$, such that

$$\boldsymbol{y}_0 = \boldsymbol{\phi}(\boldsymbol{x}_0) \qquad \Leftrightarrow \qquad \boldsymbol{f}\big(\boldsymbol{x}_0, \boldsymbol{\phi}(\boldsymbol{x}_0)\big) = 0 \quad \forall \boldsymbol{x} \in \mathcal{U}, \boldsymbol{\phi}(\boldsymbol{x}) \in \mathcal{V}$$

Moreover

$$\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{x}} = -\left[\frac{\partial \boldsymbol{f}\big(\boldsymbol{x}, \boldsymbol{\phi}(\boldsymbol{x})\big)}{\partial \boldsymbol{x}}\right]^{-1} \frac{\partial \boldsymbol{f}\big(\boldsymbol{x}, \boldsymbol{\phi}(\boldsymbol{x})\big)}{\partial \boldsymbol{x}} \tag{6.21}$$

General form The same theorem can be stated similarly considering an initial function $\boldsymbol{\phi}(\boldsymbol{z}) : \mathbb{R}^n \to \mathbb{R}^m$ where $n > m$. The vector $\boldsymbol{z} = (x_1, x_2, \ldots, x_n)$ can be partitioned, upon reordering of the variables, in 2 sets $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y})$ (where $\boldsymbol{x} \in \mathbb{R}^{n-m}$ contains the independent variables while $\boldsymbol{y} \in \mathbb{R}^m$ the dependent one) in such a way that the matrix $\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{y}}(\boldsymbol{x}_0, \boldsymbol{y}_0)$ is non-singular: this so allows to use the implicit function theorem to determine a map $\boldsymbol{\psi}$ such that $\boldsymbol{y} = \boldsymbol{\psi}(\boldsymbol{x})$.

We can observe in fact that if the matrix $\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{z}} \in \mathbb{R}^{m \times n}$ is full rank (hence all rows of the jacobian are linearly independent), then there exists m linearly independent columns while the remaining $n - m$ are linearly dependent from the others: reordering so the independent/dependent variables allows us to distinguish $\boldsymbol{x}$ from $\boldsymbol{y}$.
For this reason the map $\boldsymbol{\psi}$ is in general not unique because there exists multiple partitioned sets $(\boldsymbol{x}, \boldsymbol{y})$ determining a non-singular jacobian.

Semi-explicit DAE Recalling a differential algebraic equation in semi-explicit form

$$\begin{cases} \boldsymbol{x}' &= \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}, t) \\ 0 &= \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y}, t) \end{cases}$$

where $\boldsymbol{x} \in \mathbb{R}^n$ are the so called *differential states*, $\boldsymbol{y} \in \mathbb{R}^m$ the *algebraic states* and the maps are $\boldsymbol{f} :\in \mathbb{R}^{n+m+1} \to \mathbb{R}^n$ and $\boldsymbol{g} :\in \mathbb{R}^{n+m+1} \to \mathbb{R}^m$. If the matrix $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{y}} \in \mathbb{R}^{m \times m}$ is non-singular, then for the implicit function theorem we can define the independent and dependent variables respectively as $\boldsymbol{z}^i = (\boldsymbol{x}, t), \boldsymbol{z}^d = \boldsymbol{y}$ such that exists a map ψ for which $\boldsymbol{y} = \psi(\boldsymbol{x}, t)$ and

$$\boldsymbol{y}' = \frac{\partial \psi}{\partial \boldsymbol{x}} \boldsymbol{x}' + \frac{\partial \psi}{\partial t} = \frac{\partial \psi}{\partial \boldsymbol{x}} f + \frac{\partial \psi}{\partial t} = \boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y}, t)$$

This means that after 1 derivation the initial DAE in explicit form can be regarded as following ODE:

$$\begin{cases} \boldsymbol{x}' &= \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}, t) \\ \boldsymbol{y}' &= \boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y}, t) \end{cases} \quad + \quad \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y}, t) = 0 : \text{hidden invariant}$$

Part III

Modelling & Simulation

Chapter 7

Introduction

Mechatronics multi-body domain systems **MBDS** are *complex* systems that interacts and embeds **multiple physical domain**, such the thermal, controls, electrical, hydraulic...
Such system are intrinsically complex and their **modelling** and consequent **simulation** requires a high level of (mathematical) abstraction and modularisation; to improve the management of the model complexity it's necessary to decompose the initial MBDS into sub-system with low interaction with each other in order to focus independently on various aspect of the complex system.

From a mathematical standpoint multi-body domain systems are usually described by **differential algebraic equations** DAE (described in part II), a combination of ordinary differential and algebraic equations that can be used, as example, to perform **kinematic analysis** (to determine position/velocity/acceleration of components in the system) or **dynamic simulation** both off-line (to estimate the behaviour of the system) or online (with the **hardware in loop** HIL method allowing to control the mechatronic system).

System modelling The **model** is the process that, starting from real world observation, construct an **abstract representation** (mainly mathematical) **validated** with physical experiments that allow to generate **simulations** in order to accomplish the **analysis** of the results obtained. The realisation of a model-based analysis of a mechatronic system's dynamic can per performed through this step:

- generate a qualitative system model in order to understand the general behaviour;
- determine the domain specific models;
- specify the mathematical dynamic of the models yet generated;
- at this stage we so have a model that can predict the behaviour of real mechatronic systems.

The modelling of the system can be performed using two types of approaches:

- the **qualitative** model is a *black box* that for a given input determines an output; implementation of this modelling are **neural networks** that firstly need to be trained in order to predict the functionality of the system. It's proven in fact that any system can be described, within a certain accuracy range, by a proper neural network; the problem of this approach is that each time a parameter changes, the neural network must be re-trained and so it doesn't allow to state general assertion regarding the functionality of the multi-body system;
- a **quantitative** approach, based on differential algebraic equations, allows instead to have a more general meaning; the system is in fact described by a set of parameters, variables, equations and constraints that can somehow be symbolically manipulated or numerically solved.

All models are so approximation of the real behaviour of the system and for that reason must be validated using experiments/testing and/or using the past experience gained; every model has a limited range of validity on which results are reliable within a certain standard. A good model must

be *simple* (it's useless to have an over-complex system to analyse every little detail) and captures the critical property of the system is modelling.

Mathematical models can be *small* or *large*, *simple* or *complicated*, static or dynamic (regarding time-variant), deterministic or stochastic, qualitative or quantitative, linear or non-linear, continuous or discrete-time...

Simulations For the modelling and consequent simulation of multi-body systems several general-purpose simulation modelling and languages have been developed that can be classified according the following criteria:

- graph or language based;
- procedural or declarative (based on equation) models;
- multi or single-domain modelling respect to specific problem;
- continuous or discrete (hybrid solutions also exists): this is related to the way on how subsystems of different domain interact together. In a **co-simulation** approach the communication interval between the system is discrete and introduces a delay between the system that can cause instability, while with a **unified** approach the solver solves the full set of equation simultaneously increasing accuracy/stability but also numerical complexity;
- functional or object oriented.

Usually a symbolic approach for the modelling is preferred because it allows to be *more general*: we don't have to specify to the model what is the input and what's the output (that's instead decided at simulation time by reversing all relations) and allows to perform mathematical simplifications; usually for fast/real-time numerical simulation analytical formulas are converted in optimized `C`, `C++` code.

Chapter 8

Kinematics

The study of **kinematics** of mechanisms are useful in order to define relative positions of bodies, reconstruct body position/orientation from sensor and optimise body position/orientation evolution over time.

To describe **pose** and **attitude** (orientation) of a body in the space is better starting of with the notation of points; in particular any point P in a dimensional space (for practical MBD application the 3D environment $\mathbf{R}^3$ is considered) can be described by a vector $\boldsymbol{v}$ that can be regarded as linear combination of the canonical base $\{\hat{\boldsymbol{e}_1}, \hat{\boldsymbol{e}_2}, \hat{\boldsymbol{e}_3}\}$:

$$\boldsymbol{v} := x\hat{\boldsymbol{e}_1} + y\hat{\boldsymbol{e}_2} + z\hat{\boldsymbol{e}_3} \tag{8.1}$$

To describe a body, determining it's **configuration**, we can use a set of parameter such specific points of the body itself or some vectors on it.

Rigid bodies A **body** is defined as **rigid** if the distance between any it's two point and the angle of any two vector are constant in time (or the difference is so small that can be neglected) regardless of external forces exerted on it; mathematically a body is rigid if

$$\overrightarrow{P_1P_2} = \text{constant} \quad \text{and} \quad \angle\{\overrightarrow{P_1P_2}, \overrightarrow{P_1P_3}\} = \text{constant} \qquad \forall P_1, P_2, P_3 \text{ points} \in \text{body} \tag{8.2}$$

8.1 Rotational matrix approach

The **rotational matrix** is an approach that's used to solve kinematic problems of multi-body mechanical systems.

Considering a mono-dimensional case on where we want to describe a point lying on an axis, in order to define it's position is necessary to determine a **reference frame** on which we can define the position x of the point respect to such origin; figure 8.1 shows that different reference frames can result in different coordinates of the point.

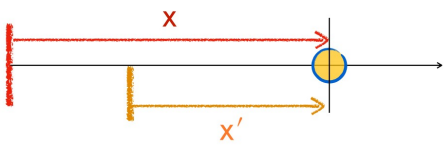


Figure 8.1: representation of the position of a point moving on one axis using two different reference frames (red and orange).

Extending the case of a planar body, it's **configuration** q can be described by defining a position (two *spatial* coordinates x, y) and orientation (θ) respect to a reference frame. Bodies are described using local reference frames where for example a point $\boldsymbol{P}_1^b$ is described. In order to define the position $\boldsymbol{P}_1$ in the global reference frame is so mandatory to know the configuration $\boldsymbol{q} = (x_0, y_0, \theta)$ of the body respect to the global coordinates (as shown in figure 8.2).

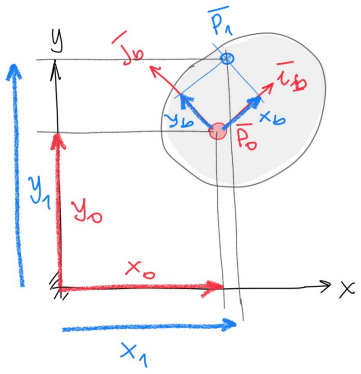


Figure 8.2: reference frames used for determining the position of a point/vector in the "global" frame knowing it's local coordinates.

In this case we can see that the **point** can be described in the global reference frame as $\boldsymbol{P}_1 = (x_1, y_1)$ knowing it's local coordinates $\boldsymbol{P}_b(x_b, y_b)$ and it's configuration $\boldsymbol{q} = (\boldsymbol{P}_0, \theta)$ as

$$\begin{aligned} \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} &= \begin{pmatrix} x_0 + x_b\cos\theta - y_b\sin\theta \\ y_0 + x_b\sin\theta + y_b\cos\theta \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} + \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{pmatrix} x_b \\ y_b \end{pmatrix} \\ \boldsymbol{P}_1 &= \boldsymbol{P}_0 + \mathcal{R}\boldsymbol{P}^b \end{aligned} \tag{8.3}$$

If we now define $\overrightarrow{P_1P_0} = \hat{\boldsymbol{\imath}}_b$ the versor (a vector such that $\|\hat{\boldsymbol{\imath}}_b\| = 1$) aligned to the x coordinate axis, we have that it description in the global reference system can be computed by determining $\boldsymbol{P}_1 - \boldsymbol{P}_0$ resulting in

$$\hat{\boldsymbol{\imath}}_b = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}$$

If we instead would have considered the versor $\hat{\boldsymbol{\jmath}}_b$ aligned to the y axis what we would have obtained is the vector $(-\sin\theta, \cos\theta)$. We can so now see that $\hat{\boldsymbol{\imath}}_b$ and $\hat{\boldsymbol{\jmath}}_b$ are representing the columns of the **rotation matrix** $\mathcal{R}$ firstly shown in equation 8.3. This matrix, in a more general case, contains the expression of the verso of the local frame measured in the global one and they generate a **base**, a set of vector $\{\hat{\boldsymbol{e}}_1, \hat{\boldsymbol{e}}_2, \hat{\boldsymbol{e}}_3\}$ characterized by having:

$$\hat{\boldsymbol{e}}_i \cdot \hat{\boldsymbol{e}}_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \qquad \text{and} \qquad \hat{\boldsymbol{e}}_1 \cdot (\hat{\boldsymbol{e}}_2 \times \hat{\boldsymbol{e}}_3) = 1$$

where the second condition represent the *right-hand rule*. We so define **world/ground reference frame** as the fixed one on respect with **local/moving** reference frames are described (and usually are attached to bodies in the system). Given so the **rotational matrix** $\mathcal{R}$ that describes the rotation of the local system respect to the world frame and the origin $\boldsymbol{x}_0$ of the local system (respect to ground) we have that the coordinates of point in the moving reference system $\boldsymbol{x}^b$ relates to the absolute space position $\boldsymbol{x}^w$ using equation

$$\boldsymbol{x}^w = \boldsymbol{x}_0 + \mathcal{R}\boldsymbol{x}^b \tag{8.4}$$

Such equation can be inverted to determining

$$\boldsymbol{x}^b = \mathcal{R}^{-1}(\boldsymbol{x}^w - \boldsymbol{x}_0) = \mathcal{R}^{-1}\boldsymbol{x}^w - \mathcal{R}^{-1}\boldsymbol{x}_0 \tag{8.5}$$

Inverse rotation Rotational matrix $\mathcal{R}$ belongs to the **special orthogonal matrix** $SO(N)$ space characterized by having

$$\det\mathcal{R} = 1 \qquad \text{and} \qquad \mathcal{R}^{-1} = \mathcal{R}^t \quad \Rightarrow \quad \mathcal{R}^{-1}\mathcal{R} = \mathcal{R}^t\mathcal{R} = \mathcal{I} \tag{8.6}$$

where so the inverse corresponds to the transposed of the matrix: this consideration is very useful because it allows to better perform inverse operations.

8.1.1 Transformation matrix

A good way to describe roto-translation as shown in equation 8.4 and 8.5 is by using the **transformation matrix** $\mathfrak{R}$ notation, where all the calculation are condensed in the 4×4 matrix described as

$$
\begin{aligned}
8.4 \mapsto \begin{pmatrix} x_w \\ y_w \\ z_w \\ 1 \end{pmatrix} &= \left[\begin{array}{ccc|c} & \mathcal{R} & & \begin{matrix} x_0 \\ y_0 \\ z_0 \end{matrix} \\ \hline 0 & 0 & 0 & 1 \end{array}\right] \begin{pmatrix} x_b \\ y_b \\ z_b \\ 1 \end{pmatrix} \\
8.5 \mapsto \begin{pmatrix} x_b \\ y_b \\ z_b \\ 1 \end{pmatrix} &= \left[\begin{array}{ccc|c} & \mathcal{R}^{-1} & & -\mathcal{R}^{-1}\boldsymbol{O}^w \\ \hline 0 & 0 & 0 & 1 \end{array}\right] \begin{pmatrix} x_w \\ y_w \\ z_w \\ 1 \end{pmatrix}
\end{aligned}
\tag{8.7}
$$

where $\boldsymbol{O}^w = (x_0, y_0, z_0)$ is the origin of the local reference frame respect to ground. With this definition the **reference frame** is the one characterized by having a transformation matrix of the form

$$
\mathfrak{R}^w = \left[\begin{array}{ccc|c} & \mathcal{I} & & \begin{matrix} 0 \\ 0 \\ 0 \end{matrix} \\ \hline 0 & 0 & 0 & 1 \end{array}\right]
$$

Operation between points and vectors While performing operation with points and/or vectors it necessary to understand if such operation is feasible (having a *physical meaning*) or not; in particular we have that

$$
\begin{aligned}
\text{point} - \text{point} &\mapsto \text{vector} \\
\text{point} + \text{vector} &\mapsto \text{point} \\
\text{vector} \pm \text{vector} &\mapsto \text{vector} \\
\text{point} + \text{point} &\mapsto \text{nothing}
\end{aligned}
$$

Order of transformation

Local frames of multi-body systems are usually realised by performing multiple *recursive* roto-traslation of reference systems. An important think to remember that the application of transformation matrix is a non-commutative operation, meaning that given the reference frame $\mathfrak{R}^w$ and any two transformation matrix $\mathfrak{R}_1$ and $\mathfrak{R}_2$ we have that

$$
\mathfrak{R}_1\mathfrak{R}_2 \neq \mathfrak{R}_2\mathfrak{R}_1
$$

Observe that the product of any 2 (or more) reference frames $\mathfrak{R}_i\mathfrak{R}_j$ generates a new transformation matrix that allow to relates this new reference frames to the global coordinates.

Pure translation If we consider a reference frame $\mathfrak{R}_2$ defined as pure translation of a frame $\mathfrak{R}_1$ (that's also a pure translation of a world reference frame $\mathfrak{R}^w$), as shown in figure 8.3, defined by the transformation matrices

$$
\mathfrak{R}_1 = \left[\begin{array}{ccc|c} & \mathcal{I} & & \begin{matrix} x_{01}^w \\ y_{01}^w \\ z_{01}^w \end{matrix} \\ \hline 0 & 0 & 0 & 1 \end{array}\right] \qquad \mathfrak{R}_2 = \left[\begin{array}{ccc|c} & \mathcal{I} & & \begin{matrix} x'_{02} \\ y'_{02} \\ z'_{02} \end{matrix} \\ \hline 0 & 0 & 0 & 1 \end{array}\right]
$$

in order to describe the second reference frame into global coordinate system we have to perform the operation $\mathfrak{R}_1\mathfrak{R}_2$. Performing algebraically the operation we indeed retrieve the intuitive result of a

transformation matrix $\mathfrak{R}_2^w$ with no rotation ($\mathcal{R} = \mathcal{I}$) and center of the base in $\boldsymbol{x}_1^w + \boldsymbol{x}_2'$, in fact

$$\mathfrak{R}_2^w = \mathfrak{R}_1\mathfrak{R}_2 = \left[\begin{array}{ccc|c} & \mathcal{I} & & \begin{matrix} x_{01}^2 + x_{02}' \\ y_{01}^2 + y_{02}' \\ z_{01}^2 + z_{02}' \end{matrix} \\ \hline 0 & 0 & 0 & 1 \end{array}\right]$$

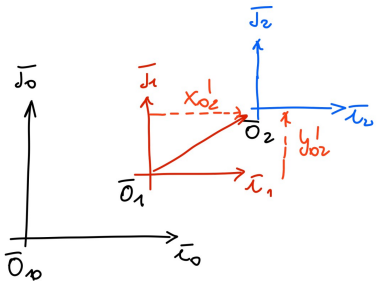


Figure 8.3: multiple transformations of pure translation; in this case, for sake of simplicity, the planar case has been considered.

Pure rotation Considering a reference frame $\mathfrak{R}_2$ that a pure rotation (along the z axis) of an angle β respect to a reference frame $\mathfrak{R}_1$ characterized by a pure rotation of angle α respect to the reference frame $\mathfrak{R}^w$ (figure 8.4), their transformation matrices are

$$\mathfrak{R}_1 = \left[\begin{array}{ccc|c} \cos\alpha & -\sin\alpha & 0 & 0 \\ \sin\alpha & \cos\alpha & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right] \qquad \mathfrak{R}_2 = \left[\begin{array}{ccc|c} \cos\beta & -\sin\beta & 0 & 0 \\ \sin\beta & \cos\beta & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right]$$

In order to determine the transformation of the second reference frame respect to the ground we so have to compute the product $\mathfrak{R}_1\mathfrak{R}_2$ between the transformation matrix, hence

$$\mathfrak{R}_2^w = \mathfrak{R}_1\mathfrak{R}_2 = \left[\begin{array}{ccc|c} \cos\alpha\cos\beta - \sin\alpha\sin\beta & -\cos\alpha\sin\beta - \sin\alpha\cos\beta & 0 & 0 \\ \sin\alpha\cos\beta + \cos\alpha\sin\beta & -\sin\alpha\sin\beta + \cos\alpha\cos\beta & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right]$$

Using so Prostaferesi's equation involving the sum of the argument in (co)sine functions we obtain the intuitive result of a pure revolution of $\alpha + \beta$ along the z axis:

$$\mathfrak{R}_2^w = \left[\begin{array}{ccc|c} \cos(\alpha+\beta) & -\sin(\alpha+\beta) & 0 & 0 \\ \sin(\alpha+\beta) & \cos(\alpha+\beta) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right]$$

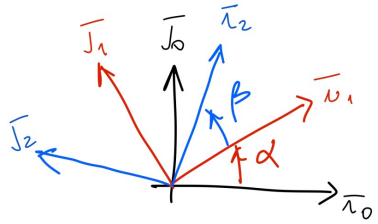


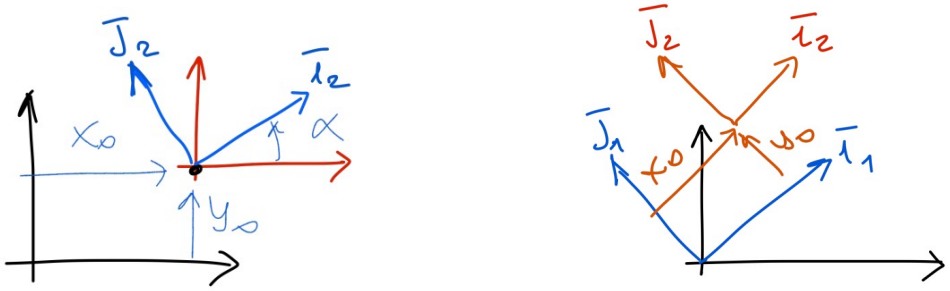
Figure 8.4: multiple transformations of pure rotation revolving the z axis.

Rotation and translation In the case of pure rotation/translation the reference frame that we have obtained was the same if we would have applied the transformation in reversed order, however this are only particular case. If we consider a translation $\mathfrak{R}_1$ and a rotation $\mathfrak{R}_2$ defined by matrices

$$\mathfrak{R}_1 = \left[\begin{array}{ccc|c} & \mathcal{I} & & \begin{matrix} x_{01}^w \\ y_{01}^w \\ 0 \end{matrix} \\ \hline 0 & 0 & 0 & 1 \end{array}\right] \qquad \mathfrak{R}_2 = \left[\begin{array}{ccc|c} \cos\alpha & -\sin\alpha & 0 & 0 \\ \sin\alpha & \cos\alpha & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right]$$

Intuitively the reference frame obtain by firstly applying the translation ($\mathfrak{R}_1$) is different from the one obtained by rotating first ($\mathfrak{R}_2$), as also shown in figure 8.5, in fact by performing the matrix calculations we obtain

$$\mathfrak{R}_1\mathfrak{R}_2 = \left[\begin{array}{ccc|c} & \mathcal{R}_\alpha & & \begin{matrix} x_{01}^w \\ y_{01}^w \\ 0 \end{matrix} \\ \hline 0 & 0 & 0 & 1 \end{array}\right] \qquad \mathfrak{R}_2\mathfrak{R}_1 = \left[\begin{array}{ccc|c} & \mathcal{R}_\alpha & & \begin{matrix} x_{01}^w \cos\alpha - y_{01}^2 \sin\alpha \\ x_{01}^2 \sin\alpha + y_{01}^w \cos\alpha \\ 0 \end{matrix} \\ \hline 0 & 0 & 0 & 1 \end{array}\right]$$



(a) (b)

Figure 8.5: reference frame obtained by first translating and than rotating (a) and first rotating and then translating (b).

8.1.2 Primitive rotation matrices

We can build any **rotation matrix** $\mathcal{R}$ by using **3 independent sequences of rotation**; they have to be 3 since that's the minimum number of independent coordinates for describing the attitude of a body and they also must be independent since we cannot repeat the rotation around the same axis unless there is an intermediate one.

Naming $\mathcal{R}_x, \mathcal{R}_y, \mathcal{R}_z$ the rotation matrix respect to axis x, y, z respectively, examples of independent (hence allowed) rotation transformation are described by the products

$$\mathcal{R}_x\mathcal{R}_y\mathcal{R}_z \qquad \mathcal{R}_y\mathcal{R}_z\mathcal{R}_x \qquad \mathcal{R}_x\mathcal{R}_y\mathcal{R}_x$$

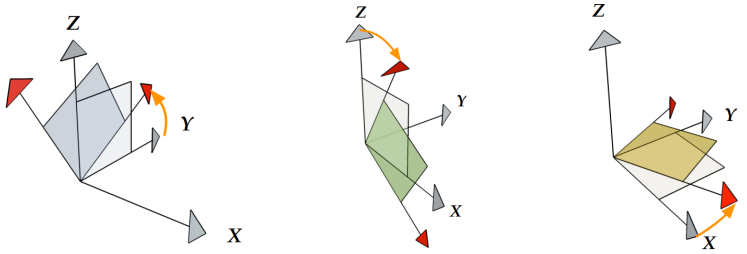


Figure 8.6: rotation matrices respect to the x (left), y (middle) and z (right) axis.

As shown in the last example we can use the same rotation matrix along the same axis if between them a different rotation matrix is used; if we would have considered a rotation $\mathcal{R}_z\mathcal{R}_z\mathcal{R}_x$ we would have had two dependent rotation $\mathcal{R}_z\mathcal{R}_z$ around the z axis that provided only one information regarding the attitude of the system (and not 2).

CHIEDERE EULER AND BRYANT ANGLES

Usually given a transformation around three consequent axis (as example (z, x, z)) with angles ϕ, θ, ψ, then other sequences (such (x, y, z)) angles can be computed from them, and so choosing different order of rotation axis leads to different rotation angles that can still describe the same attitude (so the same reference frame).

Inverse matrix Considering a direct sequence that determines the attitude of a body, in order to determine the inverse sequence we can simply reversing the rotation sequence and changing the sign of the angles. An example is determined by

$$\text{direct seq.: } \mathcal{R} = \mathcal{R}_z(\psi)\mathcal{R}_x(\phi)\mathcal{R}_y(\theta) \qquad \mapsto \qquad \text{inverse seq.: } \mathcal{R}^{-1} = \mathcal{R}_y(-\theta)\mathcal{R}_x(-\phi)\mathcal{R}_z(-\psi)$$

Singular configuration All sequences of rotation generate a rotation matrix with a **singular configuration** (also known as *Gimbal lock*) where we cannot distinguish a rotation from another.
This occurs when the rotational axis of the middle term in the sequence becomes parallel to the rotation axis of the first and third term. The loss of a degree of freedom means that mathematically we cannot invert the transformation, we can only establish a linear relationship between two of the angles; in this case the best we can do is determine the sum of *pitch* and *yaw* angles.

Considering as example the sequence of rotation z, x, y, we have that the singularity occurs for an angle around x equal to $\pi/2$, in fact it can be shown that

$$\mathcal{R}_x(\psi)\mathcal{R}_x(\pi/2)\mathcal{R}_y(\phi) = \begin{bmatrix} \cos(\psi+\phi) & 0 & \sin(\psi+\phi) \\ \sin(\psi+\phi) & 0 & -\cos(\psi+\phi) \\ 0 & 1 & 0 \end{bmatrix}$$

A singular configuration is obtained for the sequence z, x, z and a null angle considering the transformation $\mathcal{R}_z(\psi)\mathcal{R}_x(0)\mathcal{R}_z(\phi)$.

From rotation matrix to angle

Generally what we have is the *final* rotation matrix $\mathbf{R}$ where the versor of the local base are described in the base of it's reference frame, hence in the form

$$\mathcal{R} = \begin{bmatrix} u_x & v_x & w_x \\ u_y & v_y & w_y \\ u_z & v_z & w_z \end{bmatrix} \tag{8.8}$$

This matrix contains 9 parameters and so the problem is determine the values of the angles associated to a set of pure rotation that allow to determine the same attitude in space, such as example $\mathcal{R}_z(\phi)\mathcal{R}_x(\theta)\mathcal{R}_z(\psi)$.

MAGARI RIVEDERE

8.1.3 Rotation matrix properties

A general question that might arise if it exists a vector that's not affected by a rotation θ around it, meaning that if exists a **rotation axis**. Mathematically this is the **eigenvalue problem** states as

$$\mathcal{R}\boldsymbol{v}_b = \lambda \boldsymbol{v}_w$$

where the vector $\boldsymbol{v}_b$ in the local frame and $\boldsymbol{v}_w$ in the world reference systems are equals ($\boldsymbol{v}_b = \boldsymbol{v}_w = \boldsymbol{v}$). Reversing this equation we can formulate the eigenvalue problem as

$$(\mathcal{R} - \lambda\mathcal{I})\boldsymbol{v} = \mathbf{0}$$

for which is proven that a solution of the characteristic polynomial is the value $\lambda = 1$, hence we ensure that for any rotation matrix exists an eigenvector with unitary eigenvalue that represents the **rotation axis**. Defining $\text{tr}\{\mathcal{R}\}$ the trace of the matrix $\mathcal{R}$ (the sum of the principle diagonal elements), the other 2 eigenvalues are proven to be complex in the form

$$\lambda_{2,3} = \frac{\text{tr}\{\mathcal{R}\} - 1}{2} \pm j\sqrt{1 - \left(\frac{\text{tr}\{\mathcal{R}\} - 1}{2}\right)^2} = \cos\theta \pm j\sin\theta = e^{\pm j\theta}$$

and so we can compute the revolution angle around the rotation axis as $\theta = \arccos\left(\frac{\text{tr}\{\mathcal{R}\}-1}{2}\right)$.

Defining with $\mathcal{R}(\hat{\boldsymbol{n}}, \theta)$ the rotation about an axis $\hat{\boldsymbol{n}}$ of an angle θ, we that can observe the following property of the rotation:

$$\begin{aligned} &i) \quad \mathcal{R}(\hat{\boldsymbol{n}}, \theta) = \mathcal{R}(-\hat{\boldsymbol{n}}, -\theta) \\ &ii) \quad \mathcal{R}(\hat{\boldsymbol{n}}, \theta + 2\pi k) = \mathcal{R}(\hat{\boldsymbol{n}}, \theta) \qquad \forall k \in \mathbb{Z} \\ &iii) \quad \text{if } \theta = 0 \text{ then the rotation is undetermined} \end{aligned}$$

The goal now is to define a rotation around a generic axis $\hat{\boldsymbol{n}}$ of an angle θ respect to a given reference frame; in order to do so we can firstly apply any transformation that allows to have the third component of the vase $\hat{\boldsymbol{v}}$ parallel to the chosen direction $\hat{\boldsymbol{n}}$; we can so proceed applying the rotation along z of the angle θ and then we have to reverse the initial transformation. An example of rotation around a give axis $\hat{\boldsymbol{n}}$ is

$$\mathcal{R}(\hat{\boldsymbol{n}}, \theta) = \mathcal{R}_z(\alpha)\mathcal{R}_y(\beta)\mathcal{R}_z(\theta)\mathcal{R}_y(-\beta)\mathcal{R}_z(-\alpha) = \mathcal{R}_{\alpha,\beta}\mathcal{R}_z(\theta)\mathcal{R}_{\alpha,\beta}^{-1}$$

Note that the sequence $\mathcal{R}_{\alpha,\beta}$ is completely arbitrary and has the pure goal to create a reference frame whose direction z is aligned with $\hat{\boldsymbol{n}}$ and for that reason 2 independent rotation matrices are required in general.

Approach 1 In order to determine the values of the rotation angle α, β, a solution can be obtained computing the mutually orthogonal versor $\hat{\boldsymbol{i}}, \hat{\boldsymbol{j}}$ of the rotated base (functions of the rotation angle α, β) and impose that their scalar product with the direction $\hat{\boldsymbol{n}}$ is zero:

$$\begin{cases} \hat{\boldsymbol{n}} \cdot \hat{\boldsymbol{i}}_2 = 0 \\ \hat{\boldsymbol{n}} \cdot \hat{\boldsymbol{j}}_2 = 0 \end{cases}$$

We so have a non-linear system of 2 equation in 2 unknowns (α, β) that has 2 distinct solution depending on the direction (one solution give $\hat{\boldsymbol{n}}$, the other $-\hat{\boldsymbol{n}}$).

Approach 2 Another approach to determine a rotation matrix given an axis $\hat{\boldsymbol{n}}$ and the rotation angle θ is based on the **Rodrigues formula**. Given a vector $\boldsymbol{r}_w$ expressed in the world reference frame, it can be regarded as

$$\boldsymbol{r}_w = \mathcal{R}(\hat{\boldsymbol{n}}, \theta)\boldsymbol{r}_b \tag{8.9}$$

The vector $\boldsymbol{r}_b$ in the rotated system can be decomposed in a component $\boldsymbol{r}_b^{\parallel}$ parallel to $\hat{\boldsymbol{n}}$ and a orthogonal $\hat{\boldsymbol{r}}_b^{\perp}$ component that's the lonely affected by the rotation. We can in fact see that $\boldsymbol{r}_b^{\parallel} = (\hat{\boldsymbol{n}} \cdot \boldsymbol{r}_b)\hat{\boldsymbol{n}} = |r_b^{\parallel}|\hat{\boldsymbol{n}}$ and so the perpendicular component can be regarded as

$$\boldsymbol{r}_b^{\perp} = \boldsymbol{r}_b - \boldsymbol{r}_b^{\parallel} = \boldsymbol{r}_b - (\hat{\boldsymbol{n}} \cdot \boldsymbol{r}_b)\hat{\boldsymbol{n}}$$

By applying the transformation on such component, after some calculation we determine the **Rodrigues formula** as

$$\boldsymbol{r}_{w,rot} = \cos\theta\boldsymbol{r}_b + (1 - \cos\theta)(\hat{\boldsymbol{n}} \cdot \boldsymbol{r}_b)\hat{\boldsymbol{n}} + \sin(\hat{\boldsymbol{n}} \times \boldsymbol{r}_b) \tag{8.10}$$

$$
\begin{aligned}
8.4 \mapsto \begin{pmatrix} x_w \\ y_w \\ z_w \\ 1 \end{pmatrix} &= \left[\begin{array}{ccc|c} & \mathcal{R} & & \begin{matrix} x_0 \\ y_0 \\ z_0 \end{matrix} \\ \hline 0 & 0 & 0 & 1 \end{array}\right] \begin{pmatrix} x_b \\ y_b \\ z_b \\ 1 \end{pmatrix} \\
8.5 \mapsto \begin{pmatrix} x_b \\ y_b \\ z_b \\ 1 \end{pmatrix} &= \left[\begin{array}{ccc|c} & \mathcal{R}^{-1} & & -\mathcal{R}^{-1}\boldsymbol{O}^w \\ \hline 0 & 0 & 0 & 1 \end{array}\right] \begin{pmatrix} x_w \\ y_w \\ z_w \\ 1 \end{pmatrix}
\end{aligned}
\tag{8.11}
$$

Chapter 9

Rigid Body Kinematics and Rotation Matrix

Kinematic is the study of purely geometrical aspects of individual positions and motions of rigid bodies, hence cause of motions (forces and torques) are not considered. In particular we define **bodies** as **rigid** if for every point $\boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3$ in it the mutual distance between them and the angle represented by the vector connecting each pair of point are constant (or can be approximated as so):

$$\boldsymbol{P}_1\boldsymbol{P}_2 = \text{const.} \qquad \text{and} \qquad \angle\{\boldsymbol{P}_1\boldsymbol{P}_2, \boldsymbol{P}_1\boldsymbol{P}_3\} = \text{const.} \qquad \forall \boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3 \text{ points} \in \text{body}$$

9.1 Rotation matrix

To describe the **pose** (*position*) and **attitude** (*orientation in space*) of a body respect to another one (or respect to ground) we use the so called **reference frame** $\mathfrak{R}$ describing the **configuration** (the set of parameters that fully describes the system) of the body in the space. Considering planar systems, each body has 3 degrees of freedom hence the minimum number of parameters of the configuration is 3; in the spatial case the number increases to 6.

Typically bodies are described in **local** (or **moving**) reference frames $\mathfrak{R}^b$ that are obtained as roto-translations of a **world** (or **ground**) reference frame $\mathfrak{R}^w$. Given in fact the coordinates $\boldsymbol{x}_b$ of a point $\boldsymbol{P}$ described in the local frame of the body, the absolute position respect to the world frame can be obtained as

$$\boldsymbol{x}_w = \boldsymbol{x}_0 + \mathbf{R}\boldsymbol{x}_b \tag{9.1}$$

where $\boldsymbol{x}_0$ is the origin of the local frame respect to the world frame (representing so the **pose**) and $\mathbf{R}$ is a **rotation matrix** that describes the **attitude** of the body in ground coordinates.

Rotation matrix The rotation matrix, as stated, represent the attitude of a body respect to a given absolute reference. In order to specify such *angular orientation* what we need to do is to describe the vectorial base $\mathbf{e}^b$ of the body in terms of the ground vectorial base $\mathbf{e}^w$ as

$$\mathbf{e}^b = \mathbf{R}\mathbf{e}^w \tag{9.2}$$

In particular the nine elements of the matrix $\mathbf{R}$ are the generalized coordinates that describes the angular orientation of the body in the base $\mathbf{e}^w$. In particular if we consider that the 3 versor that generates the base $\mathbf{e}^b$ expressed in the world frame are $\hat{\boldsymbol{\imath}}_b = (i_{b,x}, i_{b,y}, i_{b,z})$, $\hat{\boldsymbol{\jmath}}_b = (j_{b,x}, j_{b,y}, j_{b,z})$, $\hat{\boldsymbol{k}}_b = (k_{b,x}, k_{b,y}, k_{b,z})$ then the matrix $\mathbf{R}$ can be regarded as

$$\mathbf{R} = \begin{bmatrix} \hat{\boldsymbol{\imath}}_b & \hat{\boldsymbol{\jmath}}_b & \hat{\boldsymbol{k}}_b \end{bmatrix} = \begin{bmatrix} i_{b,x} & j_{b,x} & k_{b,x} \\ i_{b,y} & j_{b,y} & k_{b,y} \\ i_{b,z} & j_{b,z} & k_{b,z} \end{bmatrix} \quad \text{or alternatively} \quad \begin{bmatrix} u_x & v_x & w_x \\ u_y & v_y & w_y \\ u_z & v_z & w_z \end{bmatrix} \tag{9.3}$$

An important aspect to keep in mind is that a base, to be so, must have that the scalar product between any different versor is zero and between the same direction must be unitary, mathematically

$$\hat{\boldsymbol{e}}_i \cdot \hat{\boldsymbol{e}}_j = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

This means that the rotation matrix $\mathbf{R}$, in order to be a proper base, must withstand the following six constraints described by all the possible combination i, j:

$$\sum_{k=1}^{3} r_{ik} r_{jk} = \delta_{ik} \qquad i, j \in \{1, 2, 3\} \tag{9.4}$$

A base is also characterized by being *right-handed* (the order of the versor can be described using the *right hand rule*) and mathematically this means imposing $\hat{\boldsymbol{e}}_1 \cdot (\hat{\boldsymbol{e}}_2 \times \hat{\boldsymbol{e}}_3) = 1$. Overall the rotation matrix $\boldsymbol{R}$ has **9 generalized coordinates** (elements of the matrix), but only **3** of them are **independent** because we have **6 constraints** determined by equation 9.4. This intuitively proves the fact that in order to describe the *orientation* of a rigid body in space only 3 parameters are required.

As it will be shown different (and more practical ways) to represent such rotation matrix will be present in order to have just 3 parameters (and not 9 subjected to 6 constraints, increasing the complexity of the problem).

Another important property related to the rotation matrix $\mathbf{R}$ is that it belongs to the **special orthogonal matrices** $SO(N)$ of order N (and in our case of spatial kinematic we have $N = 3$) that's characterized by having the inverse of $\mathbf{R}$ equals to it's transposed; we in fact have tat

$$\mathbf{R}^{-1}\mathbf{R} = \mathbf{R}^T\mathbf{R} = \mathbf{R}\mathbf{R}^T = \mathcal{I} \qquad \Leftrightarrow \qquad \mathbf{R}^{-1} = \mathbf{R}^T \tag{9.5}$$

By this means equation 9.1 can be inverted in order to express the coordinates $\boldsymbol{x}_b$ of a point in the body reference frame as function of the ground coordinates $\boldsymbol{x}_w$ as

$$\boldsymbol{x}_b = \mathbf{R}^{-1}(\boldsymbol{x}_w - \boldsymbol{x}_0) = \mathbf{R}^T\boldsymbol{x}_w - \mathbf{R}^T\boldsymbol{x}_0 \tag{9.6}$$

9.1.1 Transformation matrix

The **transformation matrix** $\mathfrak{R}$ is a 4×4 matrix that's used to fully describe the roto-translation transformation that has to be applied in order to determine the world coordinates $\boldsymbol{x}_w = (x_w, y_w, z_w)$ of a point knowing it's *position* in the local frame $\boldsymbol{x}_b = (x_b, y_b, z_b)$, it's pose $\boldsymbol{o}^w = (x_0, y_0, z_0)$ and orientation $\mathbf{R}$ respect to the global frame:

$$9.1 \mapsto \begin{pmatrix} x_w \\ y_w \\ z_w \\ 1 \end{pmatrix} = \left[\begin{array}{ccc|c} & \mathbf{R} & & \begin{matrix} x_0 \\ y_0 \\ z_0 \end{matrix} \\ \hline 0 & 0 & 0 & 1 \end{array}\right] \begin{pmatrix} x_b \\ y_b \\ z_b \\ 1 \end{pmatrix} \tag{9.7}$$

We can see that this construction directly determines the direct transformation described in equation 9.1, but we can also determine the transformation matrix of the inverse transformation as

$$9.6 \mapsto \begin{pmatrix} x_b \\ y_b \\ z_b \\ 1 \end{pmatrix} = \left[\begin{array}{ccc|c} & \mathbf{R}^{-1} & & -\mathbf{R}^{-1}\boldsymbol{o}^w \\ \hline 0 & 0 & 0 & 1 \end{array}\right] \begin{pmatrix} x_w \\ y_w \\ z_w \\ 1 \end{pmatrix} \tag{9.8}$$

Reference frames as sequence of transformations Given a world reference frame $\mathfrak{R}^w$ that's hence fixed, local frames associated to bodies are described in such immovable space using a combination of roto-translations. Given for example two transformation $\mathfrak{R}_1, \mathfrak{R}_2$ the order on which we perform the transformations heavily affects the final frame, and in general

$$\mathfrak{R}_1\mathfrak{R}_2 \neq \mathfrak{R}_2\mathfrak{R}_1$$

Pure translation If we consider a reference frame $\mathfrak{R}_2$ defined as pure translation of a frame $\mathfrak{R}_1$ (that's also a pure translation of a world reference frame $\mathfrak{R}^w$), as shown in figure 9.1, defined by the transformation matrices

$$\mathfrak{R}_1 = \left[\begin{array}{ccc|c} & \mathcal{I} & & \begin{matrix} x_{01}^w \\ y_{01}^w \\ z_{01}^w \end{matrix} \\ \hline 0 & 0 & 0 & 1 \end{array}\right] \qquad \mathfrak{R}_2 = \left[\begin{array}{ccc|c} & \mathcal{I} & & \begin{matrix} x'_{02} \\ y'_{02} \\ z'_{02} \end{matrix} \\ \hline 0 & 0 & 0 & 1 \end{array}\right]$$

in order to describe the second reference frame into global coordinate system we have to perform the operation $\mathfrak{R}_1\mathfrak{R}_2$. Performing algebraically the operation we indeed retrieve the intuitive result of a transformation matrix $\mathfrak{R}_2^w$ with no rotation ($\mathcal{R} = \mathcal{I}$) and center of the base in $\boldsymbol{x}_1^w + \boldsymbol{x}_2'$, in fact

$$\mathfrak{R}_2^w = \mathfrak{R}_1\mathfrak{R}_2 = \left[\begin{array}{ccc|c} & \mathcal{I} & & \begin{matrix} x_{01}^2 + x'_{02} \\ y_{01}^2 + y'_{02} \\ z_{01}^2 + z'_{02} \end{matrix} \\ \hline 0 & 0 & 0 & 1 \end{array}\right]$$

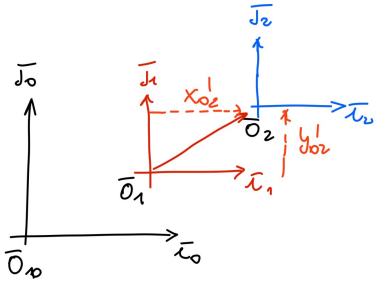


Figure 9.1: multiple transformations of pure translation; in this case, for sake of simplicity, the planar case has been considered.

Pure rotation Considering a reference frame $\mathfrak{R}_2$ that a pure rotation (along the z axis) of an angle β respect to a reference frame $\mathfrak{R}_1$ characterized by a pure rotation of angle α respect to the reference frame $\mathfrak{R}^w$ (figure 9.2), their transformation matrices are

$$\mathfrak{R}_1 = \left[\begin{array}{ccc|c} \cos\alpha & -\sin\alpha & 0 & 0 \\ \sin\alpha & \cos\alpha & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right] \qquad \mathfrak{R}_2 = \left[\begin{array}{ccc|c} \cos\beta & -\sin\beta & 0 & 0 \\ \sin\beta & \cos\beta & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right]$$

In order to determine the transformation of the second reference frame respect to the ground we so have to compute the product $\mathfrak{R}_1\mathfrak{R}_2$ between the transformation matrix, hence

$$\mathfrak{R}_2^w = \mathfrak{R}_1\mathfrak{R}_2 = \left[\begin{array}{ccc|c} \cos\alpha\cos\beta - \sin\alpha\sin\beta & -\cos\alpha\sin\beta - \sin\alpha\cos\beta & 0 & 0 \\ \sin\alpha\cos\beta + \cos\alpha\sin\beta & -\sin\alpha\sin\beta + \cos\alpha\cos\beta & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right]$$

Using so Prostaferesi's equation involving the sum of the argument in (co)sine functions we obtain the intuitive result of a pure revolution of $\alpha + \beta$ along the z axis:

$$\mathfrak{R}_2^w = \left[\begin{array}{ccc|c} \cos(\alpha+\beta) & -\sin(\alpha+\beta) & 0 & 0 \\ \sin(\alpha+\beta) & \cos(\alpha+\beta) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right]$$

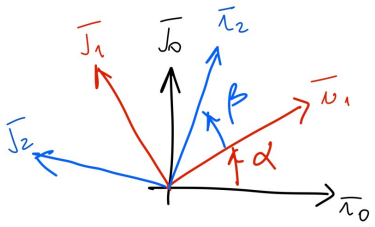


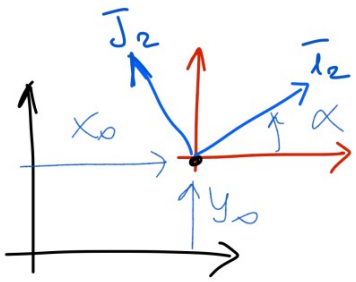
Figure 9.2: multiple transformations of pure rotation revolving the z axis.

Rotation and translation In the case of pure rotation/translation the reference frame that we have obtained was the same if we would have applied the transformation in reversed order, however this are only particular case. If we consider a translation $\mathfrak{R}_1$ and a rotation $\mathfrak{R}_2$ defined by matrices

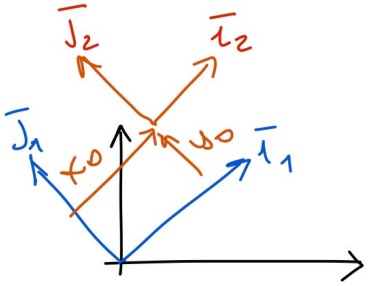
$$\mathfrak{R}_1 = \left[\begin{array}{ccc|c} & \mathcal{I} & & \begin{matrix} x_{01}^w \\ y_{01}^w \\ 0 \end{matrix} \\ \hline 0 & 0 & 0 & 1 \end{array}\right] \qquad \mathfrak{R}_2 = \left[\begin{array}{cccc} \cos\alpha & -\sin\alpha & 0 & 0 \\ \sin\alpha & \cos\alpha & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right]$$

Intuitively the reference frame obtain by firstly applying the translation ($\mathfrak{R}_1$) is different from the one obtained by rotating first ($\mathfrak{R}_2$), as also shown in figure 9.3, in fact by performing the matrix calculations we obtain

$$\mathfrak{R}_1\mathfrak{R}_2 = \left[\begin{array}{ccc|c} & \mathcal{R}_\alpha & & \begin{matrix} x_{01}^w \\ y_{01}^w \\ 0 \end{matrix} \\ \hline 0 & 0 & 0 & 1 \end{array}\right] \qquad \mathfrak{R}_2\mathfrak{R}_1 = \left[\begin{array}{ccc|c} & \mathcal{R}_\alpha & & \begin{matrix} x_{01}^w \cos\alpha - y_{01}^2 \sin\alpha \\ x_{01}^2 \sin\alpha + y_{01}^w \cos\alpha \\ 0 \end{matrix} \\ \hline 0 & 0 & 0 & 1 \end{array}\right]$$



(a)



(b)

Figure 9.3: reference frame obtained by first translating and than rotating (a) and first rotating and then translating (b).

9.1.2 Primitive rotation matrices and sequences of rotations

As stated previously, the rotation matrix $\mathbf{R}$ (equation 9.3) contains 9 parameters 3 of which are independent (the other 6 are constrained by the conditions on the mutual orthogonality) and so a *easier* formulation can be achieved using only 3 parameters. In this sense the angular orientation of the vectorial base $\mathbf{e}^b$ can be though as the result of 3 successive rotation along axis; considering as example the sequence

$$\mathbf{e}^b = \underbrace{\mathbf{R}_1(\theta)\mathbf{R}_2(\phi)\mathbf{R}_3(\psi)}_{\mathbf{R}}\,\mathbf{e}^w \qquad \leftrightarrow \qquad \mathbf{e}^b = \mathbf{R}_x(\theta)\mathbf{R}_y(\phi)\mathbf{R}_z(\psi)\mathbf{e}^w$$

a local frame can be obtained by firstly rotating the global frame along the third (z) axis by an angle ψ. This operation determines a local frame $\mathbf{e}^{b''}$ on top of which a second rotation of angle ϕ along

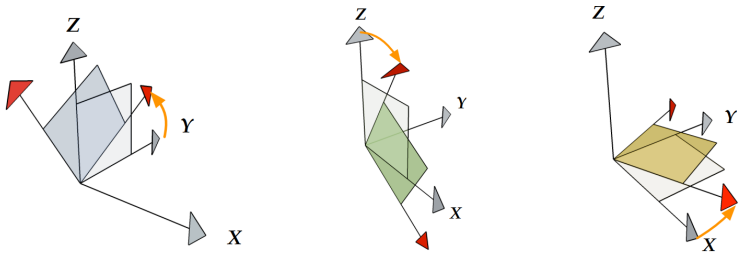


Figure 9.4: rotation matrices respect to the x (left), y (middle) and z (right) axis.

the second (y) axis is performed. That generates another vectorial space $\mathbf{e}^{b'}$ that finally determines the local base $\mathbf{e}^b$ as a rotation of an angle θ along the first (x) axis.

In particular the **primitive rotation matrices** are expressed as

$$\mathbf{R}_x(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix} \quad \mathbf{R}_y(\theta) = \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix} \quad \mathbf{R}_z(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{9.9}$$

Euler angles The **Euler angles** are a common choice in the description of the orientation in a body in space and is obtained by the sequence $z(\psi) \to x(\theta) \to z(\phi)$ (or *in numbers* by $3 \to 1 \to 3$) resulting in the transformation

$$\mathbf{R}_z(\phi)\mathbf{R}_x(\theta)\mathbf{R}_z(\psi) = \begin{bmatrix} \cos\psi\cos\phi - \sin\psi\cos\theta\sin\phi & \sin\psi\cos\theta + \cos\psi\cos\theta\sin\phi & \sin\theta\sin\phi \\ -\cos\psi\sin\phi - \sin\psi\cos\theta\cos\phi & -\sin\psi\sin\phi + \cos\psi\cos\theta\cos\phi & \sin\theta\cos\phi \\ \sin\psi\sin\theta & -\cos\psi\sin\theta & \cos\theta \end{bmatrix} \tag{9.10}$$

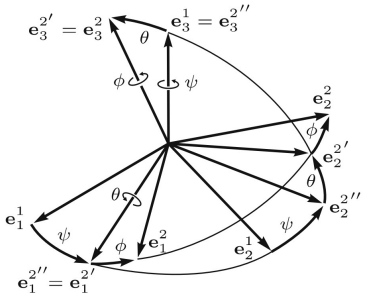


Figure 9.5: representation of the Euler angles ψ, θ, ϕ for representing the attitude of a body.

Bryan angles **Bryan** (or **Cardan**) **angles** are associated to another sequence of relative rotations carried along the sequence $1 \to 2 \to 3$ (x, y, z) resulting in

$$\mathbf{R}_z(\phi_3)\mathbf{R}_y(\phi_2)\mathbf{R}_x(\phi_1) = \begin{bmatrix} \cos\phi_2\cos\phi_3 & \cos\phi_1\sin\phi_3 + \sin\phi_1\sin\phi_2\cos\phi_3 & \sin\phi_1\sin\phi_3 - \cos\phi_1\sin\phi_2\cos\phi_3 \\ -\cos\phi_2\sin\phi_3 & \cos\phi_1\cos\phi_3 - \sin\phi_1\sin\phi_2\sin\phi_3 & \sin\phi_1\cos\phi_3 + \cos\phi_1\sin\phi_2\sin\phi_3 \\ \sin\phi_2 & -\sin\phi_1\cos\phi_2 & \cos\phi_1\cos\phi_2 \end{bmatrix} \tag{9.11}$$

Inverse sequence Until now we considered the **direct sequence** in order to describe the attitude of a local reference frame $\mathbf{e}^b$ respect to the global base $\mathbf{e}^w$, but sometime the inverse operation is required (to describe the attitude of a vector in $\mathbf{e}^b$ knowing it's coordinates in $\mathbf{e}^w$). In this case we have that the **inverse sequence** of rotation is obtained by inverting the rotation sequence and changing the sign to the angles as in this example:

$$\text{direct seq.: } \mathbf{R} = \mathbf{R}_z(\psi)\mathbf{R}_x(\phi)\mathbf{R}_y(\theta) \quad \mapsto \quad \text{inverse seq.: } \mathbf{R}^{-1} = \mathbf{R}_y(-\theta)\mathbf{R}_x(-\phi)\mathbf{R}_z(-\psi) \tag{9.12}$$

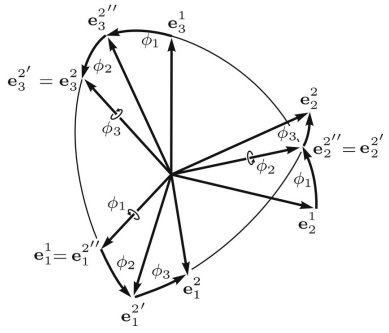


Figure 9.6: representation of the Bryan angles ϕ_1, ϕ_2, ϕ_3 for representing the attitude of a body.

Singular configuration In general the same orientation of the body can be expressed as function of different angles depending on the rotation sequence chosen, but however anyone of this representation presents a **singular configuration** (also referred s the ***gimbal lock***) where for particular values of the intermediate rotation this description doesn't allow to fully discriminate the attitude of the body. Considering as example the Euler angle representation (equation 9.10) in case of $\theta = k\pi$ (where $k \in \mathbb{Z}$), the evaluation and subsequence simplification of the matrix results in

$$\mathbf{R}_z(\phi)\mathbf{R}_x(0)\mathbf{R}_z(\psi) = \begin{bmatrix} \cos(\phi+\psi) & \sin(\phi+\psi) & 0 \\ -\sin(\phi+\psi) & \cos(\phi+\psi) & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

As we can see such sequence results in a rotation matrix along the z axis of the starting reference frame of an angle $\phi + \psi$, hence we have only one parameter (instead of the required 3) to describe the orientation of the body, hence we have 2 degree of freedom left unknown.

From rotation matrix to angles In general problems what we might have is a rotation matrix $\mathbf{R}$ (whose components are described in equation 9.3) with 9 entries subjected to 6 constraints but that we want are a sequence of primitive rotations hence their associated angles. This operation can be obtained by solving the non-linear systems associated to $\mathbf{R}_i(\phi)\mathbf{R}_j(\theta)\mathbf{R}_k(\psi) = \mathbf{R}$ (where i, j, k are generically axis of rotation); such operation should be performed on the 3 *easier* expression. Considering as example the sequence of rotation $z(\psi) \to y(\theta) \to x(\phi)$ what we obtain is the equality

$$\begin{bmatrix} \cos\theta\cos\psi & -\cos\theta\sin\psi & \sin\psi \\ \sin\phi\sin\theta\cos\psi + \cos\phi\sin\psi & -\sin\phi\sin\theta\sin\psi + \cos\phi\cos\psi & -\sin\phi\cos\theta \\ -\cos\phi\sin\theta\cos\psi + \sin\phi\sin\psi & \cos\phi\sin\theta\sin\psi + \sin\phi\cos\psi & \cos\phi\cos\theta \end{bmatrix} = \begin{bmatrix} u_x & v_x & w_x \\ u_y & v_y & w_y \\ u_z & v_z & w_z \end{bmatrix}$$

The easier terms to match are $\cos\theta\cos\phi = u_x$, $-\cos\theta\sin\psi = v_x$ and $-\sin\phi\cos\theta = w_y$ that, simultaneously solved, determines the angles

$$\begin{aligned} \phi &= \arctan\left(-\frac{u_x}{\sqrt{u_x^2 + w_y^2}}, \frac{w_y}{\sqrt{u_x^2 + w_y^2}}\right) \qquad \theta = -\arccos\left(-\sqrt{u_x^2 + w_y^2}\right) \\ \psi &= \pi - \arcsin\left(\frac{v_x}{\sqrt{u_x^2 + w_y^2}}\right) \end{aligned} \tag{9.13}$$

In reality this is just one of the multiple solutions of the non-linear system (in this particular case symbolic solvers determine 8 different solutions for the angles that satisfy the 3 equation) hence multiple combination of angles ϕ, θ, ψ can lead to the same rotation matrix.

9.1.3 Rotation axis, Euler theorem and quaternions

As was previously stated, the rotation matrix $\mathbf{R}$ relates the relative orientation of two bases $\mathbf{e}1, \mathbf{e}2$; being that such matrix lies in the set of the special orthogonal matrices is characterised by having the

inverse equal to the transpose and has $\det \mathbf{R} = 1$. Another important property that can be obtained is by considering the **eigenvalue problem** $\mathbf{R}\boldsymbol{v}_b = \lambda \boldsymbol{v}_w$, meaning that we would like to find if there's a vector $\boldsymbol{v}$ whose orientation isn't altered after the transformation has occurred. Mathematically to compute such problem we have to first compute the eigenvalues as

$$\det\left(\mathbf{R} - \lambda \mathcal{I}\right) = p(\lambda) = 0$$

where $p(\lambda)$ is the characteristic polynomial of $\mathbf{R}$ that evaluates to

$$\begin{aligned} p(\lambda) &= -\lambda^3 + \lambda^2 \mathrm{tr}\left\{\mathbf{R}\right\} - \lambda \mathrm{tr}\left\{\mathbf{R}\right\} + \det \mathbf{R} \\ &= (\lambda - 1)\left(\lambda^2 - (\mathrm{tr}\left\{\mathbf{R}\right\} - 1)\lambda + 1\right) \end{aligned}$$

It is also proven (as shown) that such characteristic polynomial has always a unitary eigenvalue, meaning that always exists one direction of transformation that remain unaltered in both *orientation* and *elongation*: this direction is hence called **rotation axis**. Moreover the other two eigenvalues $\lambda_{2,3}$ of $p(\lambda)$ are complex conjugated and are used to determine the **rotation angle** φ of the transformation along the rotation axis:

$$\begin{aligned} \lambda_{2,3} = \frac{\mathrm{tr}\left\{\mathbf{R}\right\} - 1}{2} \pm j\sqrt{1 - \left(\frac{\mathrm{tr}\left\{\mathbf{R}\right\} - 1}{2}\right)^2} &= \cos\varphi \pm j \sin\varphi = e^{\pm j\varphi} \\ \Rightarrow \qquad \varphi = \arccos\left(\frac{\mathrm{tr}\left\{\mathbf{R}\right\} - 1}{2}\right) & \end{aligned} \tag{9.14}$$

Note that the eigensystem doesn't result in one unique solution: if we denote $\mathbf{R}(\hat{\boldsymbol{n}}, \theta)$ the rotation of an angle θ along the direction $\hat{\boldsymbol{n}}$, then the same transformation can be obtained considering the opposite axis and angle:

$$\mathbf{R}(-\hat{\boldsymbol{n}}, -\theta) = \mathbf{R}(\hat{\boldsymbol{n}}, \theta)$$

This way of defining rotations is also periodic, in the sense that we can restrict any transformation along on axis to an angle $\theta \in [0, 2\pi]$

$$\mathbf{R}(\hat{\boldsymbol{n}}, \theta + 2k\pi) = \mathbf{R}(\hat{\boldsymbol{n}}, \theta) \qquad \forall k \in \mathbb{Z}$$

Moreover if we consider a rotation with rotation angle $\theta = 0$, then such transformation is undetermined (in fact no displacement occurs). With all this premises the **Euler theorem** holds and states that *the displacement of a body-fixed base from an initial position* $\mathbf{e}^1$ *to an arbitrary final position* $\mathbf{e}^2$ *is achieved by a rotation through a certain angle about an axis which is fixed in both bases. The axis has the direction of the eigenvector associated with the eigenvalue* $\lambda = 1$ *of the direction cosine matrix* $\mathbf{R}$.

Rotation matrix given an axis and the angle In practical application we might want to have that a rotation of an angle θ occurs along a certain direction $\hat{\boldsymbol{n}}$; in order so to compute the rotation matrix $\mathbf{R}(\hat{\boldsymbol{n}}, \theta)$, a way to do so is to firstly apply two generic rotation (as example $\mathbf{R}_z(\alpha)\mathbf{R}_x(\beta)$) in order to have a reference frame $\mathfrak{R}_{temp}$ whose third base component $\hat{\boldsymbol{k}}_{temp}$ is aligned with the direction $\hat{\boldsymbol{n}}$ in the ground reference frame. On such frame we can compute the rotation of the angle θ (along the z axis) but in order to obtain the final reference frame we have to apply the inverse transformation $\mathfrak{R}_{temp}^{-1} = \mathbf{R}_x(-\alpha)\mathbf{R}_z(-\beta)$ (using the result of equation 9.12). The rotation matrix can so be regarded as

$$\mathbf{R}(\hat{\boldsymbol{n}}, \theta) = \mathbf{R}_x(\alpha)\mathbf{R}_y(\beta)\mathbf{R}_z(\theta)\mathbf{R}_y(-\beta)\mathbf{R}_x(-\alpha) \tag{9.15}$$

At this point all that's left is to determine such angles α, β in order to have the alignment of the temporary reference frame with the chosen direction $\hat{\boldsymbol{n}}$: considering that the associated rotation matrix is

$$\mathbf{R}(\alpha, \beta) = \mathbf{R}_x(\alpha)\mathbf{R}_y(\beta) = \left[\hat{\boldsymbol{i}}_2(\alpha, \beta) | \hat{\boldsymbol{j}}_2(\alpha, \beta) | \hat{\boldsymbol{k}}_2(\alpha, \beta)\right]$$

where $\hat{\boldsymbol{i}}_2, \hat{\boldsymbol{j}}_2, \hat{\boldsymbol{k}}_2$ are the versor of the temporary base $\mathbf{e}^{temp}$ in global coordinates and so in order to have that $\hat{\boldsymbol{k}}_2$ is parallel to $\hat{\boldsymbol{n}}$ we have to solve the non-linear system in α, β determined as

$$\begin{cases} \hat{\boldsymbol{n}} \cdot \hat{\boldsymbol{i}}_2 = 0 \\ \hat{\boldsymbol{n}} \cdot \hat{\boldsymbol{j}}_2 = 0 \end{cases}$$

Rodrigues formula Given a rotation axis $\hat{\boldsymbol{n}}$ and the rotation angle θ, the **Rodrigues formula** allows to compute the rotated coordinates $\boldsymbol{r}_w^{rot}$ of a vector with starting components $\boldsymbol{r}_b$:

$$\boldsymbol{r}_w^{rot} = \cos\theta\, \boldsymbol{r}_b + (1-\cos\theta)(\hat{\boldsymbol{n}}\cdot\boldsymbol{r}_b)\hat{\boldsymbol{n}} + \sin\theta\, \hat{\boldsymbol{n}}\times\boldsymbol{r}_b \tag{9.16}$$

Considering that $(\hat{\boldsymbol{n}}\cdot\boldsymbol{r}_b)\hat{\boldsymbol{n}}$ can be regarded as $\boldsymbol{r}_v + \hat{\boldsymbol{n}}\times(\hat{\boldsymbol{n}}\times\hat{\boldsymbol{r}}_b)$ but also that the vectorial product $\hat{\boldsymbol{n}}\times\boldsymbol{v}$ (with $\boldsymbol{v}\in\mathbb{R}^3$ is a generic vector) can be regarded as the following matrix product

$$\mathbf{N}_s\boldsymbol{v} := \begin{bmatrix} 0 & -n_z & n_y \\ n_z & 0 & -n_x \\ -n_y & n_x & 0 \end{bmatrix} \begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix} \tag{9.17}$$

With this we can rewrite Rodrigues formula as

$$\boldsymbol{r}_w^{rot} = \Big(I + (1-\cos\theta)\mathbf{N}_s\mathbf{N}_s + \sin\theta\mathbf{N}_s\Big)\boldsymbol{r}_b \tag{9.18}$$

> **Note:** In general any vectorial product $\boldsymbol{w}\times\boldsymbol{v}$ can be reduced to a matrix multiplication in the form $\mathbf{W}\boldsymbol{v}$, where $\mathbf{W}\in\mathbb{R}^{3\times3}$ is a skew-symmetric matrix in the form shown.

By defining $q_0 = \cos(\theta/2)$ and $\tilde{\boldsymbol{q}} = (q_1,q_2,q_3) = \hat{\boldsymbol{n}}\sin(\theta/2) = \big(n_x\sin(\theta/2), n_y\sin(\theta/2), n_z\sin(\theta/2)\big)$ what we obtain are the 4 **Euler-Rodrigues parameters**; in particular $\tilde{\boldsymbol{q}}$ is a vector lying in the rotation axis (is in fact proportional by a factor $\sin(\theta/2)$ to $\hat{\boldsymbol{n}}$). Such formulation is characterized by having that

$$\sum_{i=0}^{3} q_i^2 = 1$$

The vector $\boldsymbol{q} = (q_0,\tilde{\boldsymbol{q}}) = (q_0,q_1,q_2,q_3)\in\mathbb{R}^4$ is the so called **quaternion** and allows to give a global parametrization of the rotation matrix $\mathbf{R}$ (it is said *global* because it doesn't require the definition of intermediate reference frames to determine the rotations). We can in fact show that each rotation matrix can be expressed in quaternions as

$$\mathbf{R}(\boldsymbol{q}) = \begin{bmatrix} 2q_0^2+2q_1^2-1 & -2q_0q_3+2q_1q_2 & 2q_0q_2+2q_1q_3 \\ 2q_0q_3+2q_1q_2 & 2q_0^2+2q_2^2-1 & -2q_0q_1+2q_2q_3 \\ -2q_0q_2+2q_1q_3 & 2q_0q_1+2q_2q_3 & -2q_1^2-2q_2^2+1 \end{bmatrix} \tag{9.19}$$

Such parametrization has 4 parameters, but only 3 of them are independent and so, while describing systems in such representation, the following constraint equation must always be introduced:

$$q_0^2+q_1^2+q_2^2+q_3^2 = 1$$

The cost of adding one more parameter in the description of the system (and the related constraint equation) comes with the advantage of having a representation that's always non-singular (so it doesn't have the gimbal lock problem, as for *common* rotation sequences that are intuitively easier to understand and physically interpret).

CHIEDERE LA CONFIGURAZIONE SINGOLARE PER $\theta = 0$ o $q_0 = 0$?x

9.1.4 Velocities and acceleration

Usually reference frames moves respect to each other, and so in general is important to understand how velocities (and accelerations) in local frames relates to their respective in global coordinates. For this reason we determine the **angular rate** $\boldsymbol{\omega} = (\omega_x,\omega_y,\omega_z)$ as the vector of 3 components that represents the rate of change of a reference frame with respect to another one (typically the world reference frame).

Given a vectorial base characterized by the rotation matrix $\mathbf{R}$, compute the derivative of it's components in time t determines the rate of change $\dot{\mathbf{R}}$ of the body in the world reference frame. Considering

that $\mathbf{R}^T = \mathbf{R}^{-1}$ is the inverse rotation matrix that match the world coordinates with the body one, we have that

$$\mathbf{R}^T\dot{\mathbf{R}} = \begin{bmatrix} 0 & -\omega_z & \omega_y \\ \omega_z & 0 & -\omega_x \\ -\omega_y & \omega_x & 0 \end{bmatrix} = \mathbf{P}_\Omega \tag{9.20}$$

where $\mathbf{P}_\Omega$ is a skew-symmetric matrix usual referred as **angular operator** that's used to implement the vectorial product of $\boldsymbol{\omega}$: it happens that $\boldsymbol{\omega} \times \boldsymbol{x} = \mathbf{P}_\Omega \boldsymbol{x}$ for any vector $\boldsymbol{x} \in \mathbb{R}^3$. In particular $\boldsymbol{\omega}$ represent the projection of the rotation axis in the local reference frame; moreover if we consider $\boldsymbol{u} \in \mathbb{R}^m$ the vector of the parameters used in the description of the attitude of the body, the angular rate can be regarded as a linear combination of the rate of changes of such parameters:

$$\boldsymbol{\omega}^b = \begin{pmatrix} \omega_x \\ \omega_y \\ \omega_z \end{pmatrix} = \mathbf{E}(\boldsymbol{u})\,\dot{\boldsymbol{u}} \qquad \mathbf{E}(\boldsymbol{u}) \in \mathbb{R}^{n\times m}$$

In particular

- if we consider a set of 3 independent parameters as the 3 angles $\boldsymbol{u} = (\alpha, \beta, \theta)$ generating the rotation matrix $\mathbf{R} = \mathbf{R}_x(\alpha)\mathbf{R}_y(\beta)\mathbf{R}_z(\theta)$ we can compute the angular operator $\mathbf{P}_\Omega = \mathbf{R}^T\dot{\mathbf{R}}$ determining

$$\boldsymbol{\omega}^b = \underbrace{\begin{bmatrix} \sin\beta & 0 & 1 \\ \cos\beta\sin\theta & \cos\theta & 0 \\ \cos\beta\cos\theta & \sin\beta & 0 \end{bmatrix}}_{=\mathbf{E}(\boldsymbol{u})} \underbrace{\begin{pmatrix} \dot{\alpha} \\ \dot{\beta} \\ \dot{\theta} \end{pmatrix}}_{=\dot{\boldsymbol{u}}}$$

- if we would have instead a quaternion representation $\boldsymbol{u} = (q_0, q_1, q_2, q_3)$, considering the definition in equation 9.19 for the rotation matrix, from the computation $\mathbf{R}^T\dot{\mathbf{R}} = \mathbf{P}_\Omega$ we would have obtained

$$\boldsymbol{\omega}^b = 2\begin{bmatrix} q_3 & q_2 & -q_1 & q_0 \\ q_2 & -q_1 & -q_0 & q_3 \\ -q_1 & q_0 & q_3 & q_2 \end{bmatrix} \begin{pmatrix} \dot{q}_0 \\ \dot{q}_1 \\ \dot{q}_2 \\ \dot{q}_3 \end{pmatrix}$$

  The number of free parameters is 3 because one is constrained by the equation $q_0^2 + q_1^2 + q_2^2 + q_3^2 = 1$ and so, also in this case, we have to consider the differential of such constraint in order to univocally determine the angular velocity:

$$\frac{d}{dt}(\boldsymbol{q}\cdot\boldsymbol{q} = 1) \qquad \rightarrow \qquad 2q_0\dot{q}_0 + 2q_1\dot{q}_1 + 2q_2\dot{q}_2 + 2q_3\dot{q}_3 = 0$$

We observe so that the problem of the velocities is linear if we know the actual configuration/position of the system.

Velocity Recalling equation 9.1, the global coordinates $\boldsymbol{P}^w$ of a point described by $\boldsymbol{P}^b$ in a local reference frame is defined as

$$\boldsymbol{P}^w = \boldsymbol{P}_0 + \mathbf{R}\boldsymbol{P}^b$$

where $\boldsymbol{P}_0$ is the origin of the moving frame in the global coordinates. Differentiating in time allows to determine the velocity

$$\begin{aligned} (9.1) \quad \xrightarrow{\frac{d}{dt}} \quad \dot{\boldsymbol{P}}^w &= \dot{\boldsymbol{P}}_0 + \dot{\mathbf{R}}\boldsymbol{P}^b + \mathbf{R}\dot{\boldsymbol{P}}^b = \dot{\boldsymbol{P}}_0 + \overbrace{\mathbf{R}\mathbf{R}^T}^{=\mathcal{I}}\mathbf{R}\boldsymbol{P}^b + \mathbf{R}\dot{\boldsymbol{P}}^b \\ &= \dot{\boldsymbol{P}}_0 + \mathbf{R}\mathbf{P}_\Omega\boldsymbol{P}^b + \mathbf{R}\dot{\boldsymbol{P}}^b \end{aligned} \tag{9.21}$$

where $\dot{\boldsymbol{P}}_0$ is the contribution due to the velocity in the translation of the local frame, $\mathbf{R}\mathbf{P}_\Omega\boldsymbol{P}^b$ is due to the frame angular velocity and $\mathbf{R}\dot{\boldsymbol{P}}^b$ is an optional terms that considers the relative velocity $\dot{\boldsymbol{P}}^b$ that the point might have in the local frame. The associated **transformation matrix** is so

$$\dot{\boldsymbol{P}}^w = \left[\begin{array}{c|c} \mathbf{R}\mathbf{P}_\Omega & \dot{\boldsymbol{P}}_0 \\ \hline \mathbf{0}^t & 0 \end{array}\right]\boldsymbol{P}^b + \left[\begin{array}{c|c} \mathbf{R} & \dot{\boldsymbol{P}}_0 \\ \hline \mathbf{0}^t & 1 \end{array}\right]\dot{\boldsymbol{P}}^b$$

Acceleration Differentiating 9.21 one more time respect to time allows to compute the **acceleration** of the point P in the global reference frame as

$$\xrightarrow{\frac{d}{dt}} \qquad \ddot{\boldsymbol{P}}^w = \ddot{\boldsymbol{P}}_0 + \ddot{\mathbf{R}}\boldsymbol{P}^b + \dot{\mathbf{R}}\dot{\boldsymbol{P}}^b + \dot{\mathbf{R}}\dot{\boldsymbol{P}}^b + \mathbf{R}\ddot{\boldsymbol{P}}^b \tag{9.21}$$

Considering that $\ddot{\mathbf{R}} = \frac{d}{dt}\dot{\mathbf{R}} = \frac{d}{dt}(\mathbf{R}\mathbf{P}_\Omega) = \dot{\mathbf{R}}\mathbf{P}_\Omega + \mathbf{R}\dot{\mathbf{P}}_\Omega = \mathbf{R}\mathbf{R}^T\dot{\mathbf{R}}\mathbf{P}_\Omega + \mathbf{R}\mathbf{P}_\Omega\mathbf{P}_\Omega$ what we obtain is

$$\begin{aligned}\ddot{\boldsymbol{P}}^w &= \ddot{\boldsymbol{P}}_0 + \mathbf{R}\mathbf{P}_\Omega\mathbf{P}_\Omega\boldsymbol{P}^b + \mathbf{R}\dot{\mathbf{P}}_\Omega\boldsymbol{P}^b + 2\mathbf{R}\mathbf{P}_\Omega\dot{\boldsymbol{P}}_b + \mathbf{R}\ddot{\boldsymbol{P}}^b \\ &= \ddot{\boldsymbol{P}}_0 + \underbrace{\mathbf{R}\Big(\underbrace{\mathbf{P}_\Omega\mathbf{P}_\Omega\boldsymbol{P}^b}_{\text{centripetal acc.}} + \underbrace{\dot{\mathbf{P}}_\Omega\boldsymbol{P}^b}_{\text{tangential acc.}} + \underbrace{2\mathbf{P}_\Omega\dot{\boldsymbol{P}}^b + \ddot{\boldsymbol{P}}^b}_{\text{relative acc}}\Big)}_{\text{acceleration in the local frame}}\end{aligned} \tag{9.22}$$

9.1.5 Natural coordinates

In general rotation matrices $\mathbf{R}$ can be expressed with any set of parameters we want; the idea is that the matrix $\mathbf{R} = \begin{bmatrix} i_x & j_x & k_x \\ i_y & j_y & k_y \\ i_z & j_z & k_z \end{bmatrix}$ can be described with 9 parametric components constrained by 6 equation: this determines a more redundant formulation that however generates a new way to describe the configuration of bodies. The main idea is now to use global coordinates of points (that mechanically can be associated to joints) or vector (that can represent the direction relative translation/rotation of two elements) in order to build the rotation matrix $\mathbf{R}$.

The advantage of this approach is that we have at maximum quadratic equations that can easily solved numerically by calculators with the problem that quadratic constraints generates 2 solutions each.

Planar case with two points If we consider two points P_1, P_2 lying in the same rigid body (hence their distance L is fixed) characterized by global cartesian coordinates $(x_1, y_1), (x_2, y_2)$ respectively (for simplicity we consider a planar case), the whole system can be described with the coordinates $\boldsymbol{q} = (x_1, y_1, x_2, y_2)$. In order to determine $\mathbf{R}$ (that requires 3 independent parameters) as function of $\boldsymbol{q}$ (that actually has 4 parameters) we need to determine one constraint equation.

In order to determine the moving reference frame it is necessary firstly to chose one point as origin of such frame (in this case we consider P_1) and we have to orient one versor of the frame (in this case $\hat{\boldsymbol{i}}_1$) with the vector $\boldsymbol{P_1P_2}$ connecting the two points. Being the body rigid we have that the constraint equation is related to the length between the points that's fixed, in fact

$$\boldsymbol{P_1P_2} \cdot \boldsymbol{P_1P_2} = L^2$$

By what we just stated, the first versor $\hat{\boldsymbol{i}}_1$ of the moving frame in global coordinates can be computed as the normalization of the vector $\boldsymbol{P_1P_2}$, hence

$$\hat{\boldsymbol{i}}_1 = \frac{\boldsymbol{P_1P_2}}{L} = \begin{pmatrix} \frac{x_2 - x_1}{L} \\ \frac{y_2 - y_1}{L} \\ 0 \end{pmatrix} = \begin{pmatrix} i_{1x} \\ i_{1y} \\ i_{1z} \end{pmatrix}$$

The goal now is to determine the coordinates of the other versor $\hat{\boldsymbol{j}}_1$ that compose the rotation matrix that, in the planar case, has the form

$$\mathbf{R} = \begin{bmatrix} i_{1x} & j_{1x} & 0 \\ i_{1y} & j_{1y} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Knowing that a proper rotation matrix requires $\hat{\boldsymbol{i}}_1 \cdot \hat{\boldsymbol{j}}_1 = 0$ we obtain the relation

$$i_{1x}j_{1x} + i_{1y}j_{1y} = 0 \qquad \Rightarrow \qquad j_{1x} = -\frac{i_{1y}}{i_{1x}}j_{1y} = -\frac{y_2 - y_1}{x_2 - x_1}j_{1y}$$

Knowing that it must also hold $\hat{\boldsymbol{j}}_1 \cdot \hat{\boldsymbol{j}}_1 = 1$ we build the quadratic relation

$$j_{1x}^2 + j_{1y}^2 = j_{1y}^2 \left(1 + \frac{(y_2 - y_1)^2}{(x_2 - x_1)^2}\right) = 1 \qquad \Rightarrow \qquad \Rightarrow j_{1y} = \pm \frac{x_2 - x_1}{L}$$

We so have that the 2 possible formulation of the rotation matrix are

$$\mathbf{R} = \begin{bmatrix} \frac{x_2 - x_1}{L} & -\frac{y_2 - y_1}{L} & 0 \\ \frac{y_2 - y_1}{L} & \frac{x_2 - x_1}{L} & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \text{or} \qquad \mathbf{R} = \begin{bmatrix} \frac{x_2 - x_1}{L} & \frac{y_2 - y_1}{L} & 0 \\ \frac{y_2 - y_1}{L} & -\frac{x_2 - x_1}{L} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

that are subjected to

$$(x_2 - x_1)^2 + (y_2 - y_1)^2 = L^2$$

The problem of this formulation is that it isn't intuitive the rotation α of the reference frame respect to the global direction $\hat{\boldsymbol{i}}_0$, however this can be achieved equating $\mathbf{R}$ with the rotation matrix function of α:

$$\begin{bmatrix} \frac{x_2 - x_1}{L} & -\frac{y_2 - y_1}{L} & 0 \\ \frac{y_2 - y_1}{L} & \frac{x_2 - x_1}{L} & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \Rightarrow \quad \sin\alpha = \frac{y_2 - y_1}{2} \quad \Rightarrow \quad \alpha = \arcsin\left(\frac{y_2 - y_1}{2}\right)$$

9.2 Multi-body approach

A system characterized by n_b free bodies (in the sense that they are not constrained together) results in a

$$6n_b \text{ degrees of freedom}$$

for the system; such value indeed represents also the minimum number of components for the **generalized coordinates** $\boldsymbol{q} \in \mathbb{R}^n$ of the system itself. Note that this is the minimum value, because if we use quaternion representation for representing the transformation matrices associated to bodies, each of them has 7 different parameters (not 6, but it also introduced the associated constraint equation).

Path and trajectory It is important to distinct the **path** as a sequence of configurations (that are independent from the concept of time) and the **trajectory** as the sequence of configurations that are changing with time. In this sense a trajectory can be regarded as a parametrization in time of a path.

Independent motions Each added mechanical joint results in a mathematical constraint that reduces the degree of freedom of the system; such joints are described by the vectorial map $\boldsymbol{\Phi}(\boldsymbol{q}) = \mathbf{0} \in \mathbb{R}^m$ that models m constraint equations. This reduces the number n_i of independent generalized coordinates:

$$\boldsymbol{q} = (\boldsymbol{q}_i, \boldsymbol{q}_d)$$

where $\boldsymbol{q} \in \mathbb{R}^n$ are the *original* generalized coordinates, $\boldsymbol{q}_i \in \mathbb{R}^{n_i}$ are the independent and $\boldsymbol{q}_d \in \mathbb{R}^{n_d}$ are the dependent coordinates (of course it must hold $n_i + n_d = n$). The total number of **degrees of freedom** (representing the minimum number of independent motion to the describe the configuration of the multi-body system) of the system can so be computed as

$$DOF = n - m$$

In general the constrained degrees of freedom strictly depends on the actual position and geometry of the system that's strictly related to time, and so given the constraint map $\boldsymbol{\Phi}(\boldsymbol{q}) : \mathbb{R}^n \mapsto \mathbb{R}^m$ (where $n > m$) the actual number of constrained degrees of freedom is

$$m = \text{rank}\left\{\frac{\partial \boldsymbol{\Phi}(\boldsymbol{q})}{\partial \boldsymbol{q}}\right\} \tag{9.23}$$

where $\partial\boldsymbol{\Phi}/\partial\boldsymbol{q} \in \mathbb{R}^{m \times n}$ is the Jacobian of the map $\boldsymbol{\Phi}$. If the Jacobian has so it's maximum rank it also means that $\det \frac{\partial \boldsymbol{\Phi}}{\partial \boldsymbol{q}} \neq 0$; every time such determinant is null we are in a situation where 1 or more joints are redundant: such condition might happen locally, but if it happens permanently it is necessary to eliminate the redundant equation in order to make the solver solve the problem.

Velocity constraints If we have that the constraint map $\boldsymbol{\Phi}$ set a conditions on the generalized coordinates $\boldsymbol{q}$ to be $\boldsymbol{\Phi}(\boldsymbol{q}) = \mathbf{0}$, then it means that also the velocities $\dot{\boldsymbol{q}}$ are constrained: deriving in time $\boldsymbol{\Phi}(\boldsymbol{q}) = \mathbf{0}$ results in fact

$$\frac{\partial \boldsymbol{\Phi}(\boldsymbol{q})}{\partial \boldsymbol{q}} \dot{\boldsymbol{q}} + \frac{\partial \boldsymbol{\Phi}(\boldsymbol{q})}{\partial t} = \mathbf{0} \tag{9.24}$$

Constraint classification The constraint map $\boldsymbol{\Phi}$ can be classified respect to different aspects; if the constraint explicitly depends on time $\boldsymbol{\Phi}$ is said **rehonomous** while in the other case is said **scleronomous**, mathematically

$$\boldsymbol{\Phi}\big(\boldsymbol{q}(t), t\big) = 0\text{: rehonomous} \qquad \boldsymbol{\Phi}\big(\boldsymbol{q}(t)\big) = 0\text{: scleronomous}$$

Constraint are said **holonomic** if it involves only the actual configuration of the system, while is **non-holonomic** if it requires also velocities or higher order derivative of $\boldsymbol{q}$:

$$\boldsymbol{\Phi}\big(\boldsymbol{q}(t)\big) = 0\text{: holonomic} \qquad \boldsymbol{\Phi}\big(\boldsymbol{q}(t), \dot{\boldsymbol{q}}(t), \dots\big) = 0\text{: non-holonomic}$$

9.2.1 Topology

A **multi-body** system is a collection of (multiple) bodies, joints, forces and torques; the **topology** studies in particular how such elements are connected specially involving the use od **linear graphs**.

In order to construct such graphs (as exemplified in figure 9.7) we can follow this procedure:

1. define a reference frame for each body and assign them a node (in the next figure green);
2. define auxiliary reference frames that will be used to define constraints; such elements are still nodes in the linear graph but doesn't introduce any additional coordinate because they are fixedly dependent on the relative body frame;
3. connect all the nodes with *arrow* (formally edges) representing the constraints between nodes (reference frames).

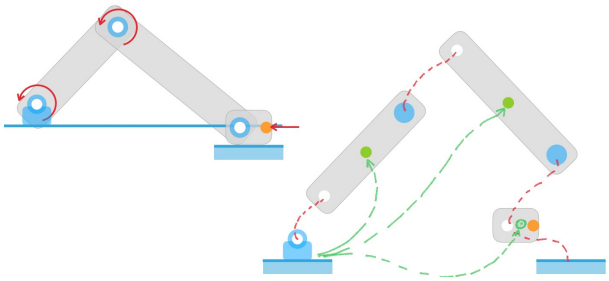


Figure 9.7: example of linear graph generation of a slider-crank mechanism where each body is characterized by a reference frame in ground coordinates (green nodes) and dashed-red lines are describing the mechanical joints.

Defining the **path** as a (finite/infinite) sequence of edges which joins a set of distinct nodes, we then have a **circuit loop** when the path begins and ends at the same vertex. In contrary a **tree** is an undirected graph in which any two vertices are connected by exactly one path.

Considering as example the linear graph in figure 9.8 we conventionally denote with h the edges associated to kinematic equation (that are strictly related to the mechanical joints), with r the edges representing fixed transformations between two reference frames in a body, with b the time-varying transformation between the body reference frame respect to ground. Auxiliary edges a might be required to describe the transformations between reference frames that are not directly connected.

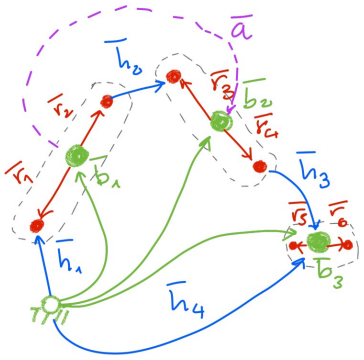


Figure 9.8: linear graph of the slider-crank mechanism of figure 9.7.

Tree and closed loops A graph is a *topological tree* if there exists exactly one path between any two nodes in the graph. If so the connectivity graph of a rigid body system is a topological tree then such system is regarded as a **kinematic tree**. Such kind of graphs have special properties that make it easy to calculate their dynamics efficiently.
Moreover a **spanning tree** of a graph G, denoted as G_t, is a sub-graph of G containing all the nodes in G together with any subset of the arcs in G such that G_t is a topological tree (as in figure 9.9); is is so proven that each connectivity graph has at least one spanning tree and if G is already a tree, then $G_t = G$ (in all other case the spanning trees are not unique).

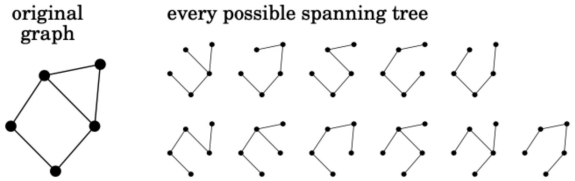


Figure 9.9: example of a graph that's not a topological tree (left) and all possible spanning trees (right).

A **cycle** is instead a closed path in a graph, hence a tree (to be so) must present no cycles. Each cycle in a connectivity graph defines a **kinematic loop**, however not all of them are useful: only the independents one must be used for the description of the system. For equation of motion only the minimal set of loops must be used. If a system has a kinematic loop then it's referred as **closed-loop system**.

9.2.2 Global and recursive approaches

Recursive formulation

In a **recursive approach** to solve the kinematic of a system we firstly need to generate a linear graph that's a tree in order so to have an open-loop system. The idea is so to determine a sequence of nodes (starting from ground for simplicity) and then follow the kinematic chain by applying all the transformation. Following such sequence we build reference frames on top of each others: each edge h is so a transformation matrix that introduces one or more generalized coordinates that are used to describe the motion of the system (hence describe the relative motion of the joint). Other edges r are instead fixed and are used to *move* inside the body.

Such method is characterized by having the minimal number of dependent coordinates; in fact if the linear graph is a tree then the number of coordinates represents the degrees of freedom of the system, while if the multi-body system has some loops we have the introduction of constraints equations.

AGGIUNGERE ESEMPIO

Global formulation

Using a **global approach** each body is initially considered as free and each body reference frame is described with parametric transformation matrices (respect to ground). Auxiliary reference frames are

constructed on top of them and then constraint equations h are applied to *assemble* the mechanism. This approach is characterized by having mainly global coordinates and simpler constraint equations, however it produces a higher number of dependent coordinates (way much more than the degrees of freedom of the system).

Appendix A

Appendix

A.1 Properties of the Laplace transform and transforms of common functions

Table A.1: useful properties of the Laplace transform and transform of common functions.

#	$f(t)$	$\mathscr{L}\{f(t)\}(s)$
1	$a\,f(t)+b\,g(t)$	$a\,\hat{f}(s)+b\,\hat{g}(s)$
2	$f(at)$	$\frac{1}{a}\hat{f}\left(\frac{s}{a}\right)$
3	$e^{at}f(t)$	$\hat{f}(s-a)$
4	$f(t-a)$	$e^{-as}\hat{f}(s)$
5	$\int_0^t f(z)\,dz$	$\frac{1}{s}\hat{f}(s)$
6	$f'(t)$	$s\hat{f}(s)-f(0^+)$
7	$f''(t)$	$s^2\hat{f}(s)-f'(0^+)-sf(0^+)$
8	$\frac{d^n}{dt^n}f(t)$	$s^n\hat{f}(s)-\sum_{j=0}^{n-1}s^{n-j-1}f^{(j)}(0^+)$
9	$t^n\,f(t)$	$(-1)^n\frac{d^n}{ds^n}\hat{f}(s)$
10	$(f\otimes g)(t)$	$\hat{f}(s)\hat{g}(s)$
1	1	$\frac{1}{s}$
2	t	$\frac{1}{s^2}$
3	t^k	$\frac{k!}{s^{k+1}}$
4	a^{bt}	$\frac{1}{s-b\log a}$
5	$e^{at}\cos(\omega t)$	$\frac{s-a}{(s-a)^2+\omega^2}$
6	$e^{at}\sin(\omega t)$	$\frac{\omega}{(s-a)^2+\omega^2}$
7	$e^{at}\cosh(\omega t)$	$\frac{s-a}{(s-a)^2-\omega^2}$
8	$e^{at}\sinh(\omega t)$	$\frac{\omega}{(s-a)^2-\omega^2}$
9	$e^{at}t^n$	$\frac{n!}{(s-a)^{n+1}}$
10	$e^{\alpha t}-e^{\beta t}$	$\frac{\alpha-\beta}{(s-\alpha)(s-\beta)}$

A.2 Resume: minimization

Given the minimization problem of the form

$$\begin{aligned} &\text{minimize:} && f(\boldsymbol{x}) && \\ &\text{subject to:} && h_k(\boldsymbol{x}) = 0 && k = 1, \dots, m \\ &&& g_k(\boldsymbol{x}) \geq 0 && k = 1, \dots, p \end{aligned}$$

the solution using the KKT can be found by firstly constructing the lagrangian $\mathcal{L}$ as

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \boldsymbol{x} - \sum_{i=1}^{m} \lambda_i h_i(\boldsymbol{x}) - \sum_{i=1}^{p} \mu_i g_i(\boldsymbol{x})$$

Candidates to be minimum point can be computed by using the first order necessary condition (stationarity of the point) that requires:

$$\begin{aligned} \nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) &= \mathbf{0} && \\ h_k(\boldsymbol{x}^*) &= 0 && k = 0, \dots, m \\ g_k(\boldsymbol{x}^*) &\geq 0 && k = 0, \dots, p \\ \mu_k^* g_k(\boldsymbol{x}^*) &= 0 && k = 0, \dots, p \\ \mu_k^* &\geq 0 && k = 0, \dots, p \end{aligned}$$

Determined the candidates, we have to compute the kernel of the gradient of the qualified constraints H (the set of linearly independent gradients of the active constraints) and verify if the hessian of the lagrangian respect to the variables $\boldsymbol{x}$ is (semi-)positive defined:

$$\boldsymbol{z}^t \nabla_{\boldsymbol{x}} \mathcal{L}\, \boldsymbol{z} \begin{cases} > 0 & \text{: sufficient condition} \\ \geq 0 & \text{: necessary condition} \end{cases} \qquad \text{with } \boldsymbol{z} \in \ker\{\nabla H\}$$

A.3 Resume: functional minimization

The minimization of a functional $\mathcal{F}(x)$ is based on determining the functions x that determines a null first variation $\delta\mathcal{F}$ (and for that reason the fundamental lemma of calculus of variation will be used).

Variation of the lagrangian Considering functionals $\mathcal{F}(x)$ that present integral relation with x, in the form

$$\mathcal{F}(x) = \int_a^b L\Big(x(t), x'(t), t\Big)\, dt$$

then the related variation obtained with integration by part is of the form

$$\delta\mathcal{F} = \int_a^b \left(\frac{\partial L}{\partial x} - \frac{d}{dt}\frac{\partial L}{\partial x'}\right)\delta_x\, dt + \left[\frac{\partial L}{\partial x'}\delta_x\right]_a^b$$

Similarly if x appears in the integral with a derivative up to the second order, then the following relation must be considered:

$$\begin{aligned}\mathcal{F}(x) &= \int_a^b L\Big(x(t), x'(t), x''(t), t\Big)\, dt \\ \delta\mathcal{F} &= \left(\frac{\partial L}{\partial x} - \frac{d}{dt}\frac{\partial L}{\partial x'} + \frac{d^2}{dt^2}\frac{\partial L}{\partial x''}\right)\delta_x\, dt + \left[\left(\frac{\partial L}{\partial x'} - \frac{d}{dt}\frac{\partial L}{\partial x''}\right)\delta_x\right]_a^b + \left[-\frac{\partial L}{\partial x''}\delta'_x\right]_a^b\end{aligned}$$

A.4 Resume: optimal control problem

Given the optimal control problem with states $\boldsymbol{x}$ and controls $\boldsymbol{u}$ in the form

$$\begin{aligned}
&\text{minimize:} && \phi\big(\boldsymbol{x}(a),\boldsymbol{x}(b)\big)+\int_a^b L(\boldsymbol{x},\boldsymbol{u},t)\,dt && \\
&\text{subejct to:} && \boldsymbol{x}'=\boldsymbol{f}(\boldsymbol{x},\boldsymbol{u},t) && \text{ODE}\\
& && \boldsymbol{B}\big(\boldsymbol{x}(a),\boldsymbol{x}(b)\big)=0 && \text{boundary conditions}\\
& && \int_a^b \boldsymbol{g}(\boldsymbol{x},\boldsymbol{u},t)\,dt=\boldsymbol{g}_0 && \text{integral constraints}\\
& && u\in\mathcal{U} && \text{control domain}
\end{aligned}$$

the solution can be obtained by firstly removing the integral constraints by replacing each one of them with new ordinary differential equation in the form $z_i'=g_i(\boldsymbol{x},\boldsymbol{u},t)$ and two other boundary condition in the form $g_i(a)=0$ and $g_i(b)=g_{i,0}$. With that stated the every optimal control problem with integral constraint can be rewritten in the form

$$\begin{aligned}
&\text{minimize:} && \phi\big(\boldsymbol{x}(a),\boldsymbol{x}(b)\big)+\int_a^b L(\boldsymbol{x},\boldsymbol{u},t)\,dt && \\
&\text{subejct to:} && \boldsymbol{x}'=\boldsymbol{f}(\boldsymbol{x},\boldsymbol{u},t) && \text{ODE}\\
& && \boldsymbol{B}\big(\boldsymbol{x}(a),\boldsymbol{x}(b)\big)=0 && \text{boundary conditions}\\
& && u\in\mathcal{U} && \text{control domain}
\end{aligned}$$

where the new added variables $\boldsymbol{z}$ associated to the integral constraints are condensed in the state variables $\boldsymbol{x}$.

With that said we can compute the hamiltonian $\mathcal{H}$ and the utility function $\mathcal{B}$ of the problem as

$$\begin{aligned}
\mathcal{H}(\boldsymbol{x},\boldsymbol{u},\boldsymbol{\lambda},t)&=L(\boldsymbol{x},\boldsymbol{u},t)+\boldsymbol{\lambda}\cdot\boldsymbol{f}(\boldsymbol{x},\boldsymbol{u},t)\\
\mathcal{B}\big(\boldsymbol{x}(a),\boldsymbol{x}(b),\boldsymbol{\mu}\big)&=\phi\big(\boldsymbol{x}(a),\boldsymbol{x}(b)\big)+\boldsymbol{\mu}\cdot\boldsymbol{B}\big(\boldsymbol{x}(a),\boldsymbol{x}(b)\big)
\end{aligned}$$

The resulting boundary valued problem of the optimal control problem is so

$$\begin{cases}
\boldsymbol{x}'=\boldsymbol{f}(x,u,t)=\dfrac{\partial\mathcal{H}}{\partial\boldsymbol{\lambda}} & \text{: original ODE}\\
\boldsymbol{\lambda}'=-\dfrac{\partial\mathcal{H}}{\partial\boldsymbol{x}} & \text{: adjoint ODE - co-equations}\\
\boldsymbol{B}\big(\boldsymbol{x}(a),\boldsymbol{x}(b)\big)=0 & \text{: original BC}\\
\left.\begin{array}{l}\dfrac{\partial\mathcal{B}}{\partial\boldsymbol{x}_a}+\boldsymbol{\lambda}(a)\\[2ex] \dfrac{\partial\mathcal{B}}{\partial\boldsymbol{x}_b}-\boldsymbol{\lambda}(b)\end{array}\right\} & \text{: adjoint BC}\\
\boldsymbol{u}(t)=\underset{\overline{\boldsymbol{u}}\in\mathcal{U}}{\operatorname{argmin}}\big\{\mathcal{H}\big(\boldsymbol{x},\overline{\boldsymbol{u}},\boldsymbol{\lambda},t\big)\big\} & \text{: Pontryagin min principle}\\
\dfrac{\partial\mathcal{H}}{\partial\boldsymbol{u}}=0 & \text{: control equation}
\end{cases}$$

where the solution of the Pontryagin minimum principle is the parametric solution that minimize the terms $\mathcal{H}$ that contains only the control $\boldsymbol{u}$ respect to the controls bound $\mathcal{U}$.

Appendix B

Final revision

B.1 January 24, 2022

Domande:

- parte A, domanda 6: perché sono sbagliate $h(x,y) = 0$ e $g(x,y) \geq 0$? Rispetto alle first order necessary condition.
- parte B, domanda 4: non mi risulta la ODE $x'' - 2x + \lambda = 1$ ma $-2x'' + 2x - \lambda + 2$
- parte B, domanda 5: non mi risulta la ODE $2x - y'' = 0$ ma $2x - z - y'' = 0$; per quanto riguarda le condizioni al contorno non mi risulta $y'(0) = y(0)$ ma $y'(0) = 0$ e neanche non mi torna $x'(0) = x(0)$;
- parte B, domanda 7: ho sbagliato io
- parte B, domanda 8: non mi tornano le adjoint boundary conditions, dovrebbe esserci solo $\lambda_2(1) = 0$ (e non $\lambda_3(0) = \lambda_3(1) = 0$)

Part B, question 4

Given the problem

$$\begin{aligned} &\text{minimize:} && \mathcal{F}(x) = x(0) + \int_0^1 \left(x^2 + \left(x' - t\right)^2\right) dt \\ &\text{subject to:} && \int_0^1 x\, dt = 0 \\ & && x(1) = 2 \end{aligned}$$

the related lagrangian is build as

$$\mathcal{L}(x, \lambda, \mu) = \int_0^1 \underbrace{\left(x^2 + \left(x' - t\right)^2 - \lambda x\right)}_{=L} dt + x(0) - \mu\big(x(1) - 2\big)$$

It's first variation can be regarded as

$$\begin{aligned} \delta\mathcal{L} &= \int_0^1 \Big(\big(2x - \lambda\big)\delta_x - 2\big(x' - t\big)\delta_{x'}\Big)\, dt - \int_0^1 x\, \delta_\lambda\, dt + \delta_{x(0)} - \mu\delta_{x(1)} - \big(x(1) - 2\big)\delta_\mu \\ &= \int_0^1 \left(\frac{\partial L}{\partial x} - \frac{d}{dt}\frac{\partial L}{\partial x'}\right)\delta_x\, dt + \left[\frac{\partial L}{\partial x'}\delta_x\right]_0^1 - \int_0^1 x\, \delta_\lambda\, dt + \delta_{x(0)} - \mu\delta_{x(1)} - \big(x(1) - 2\big)\delta_\mu \end{aligned}$$

Considering that $\frac{\partial L}{\partial x} = 2x - \lambda$, $\frac{\partial L}{\partial x'} = 2(x' - t)$ and so $\frac{d}{dt}\frac{\partial L}{\partial x'} = 2x'' - 2$ we can explicit the variation as

$$\begin{aligned}\delta\mathcal{L} = &\int_0^1 \big(2x - \lambda - 2x'' + 2\big)\delta_x\, dt + 2\big(x'(1) - 1\big)\delta_{x(1)} - 2x'(0)\delta_{x(0)} - \int_0^1 x\,\delta_\lambda\, dt \\ &+ \delta_{x(0)} - \mu\delta_{x(1)} - \big(x(1) - 2\big)\delta_\mu\end{aligned}$$

The resulting boundary value problem, solution of the functional minimization, can be obtained by setting to zero the terms related to each variation $\delta.$:

$$\begin{cases} 2x - \lambda - 2x'' + 2 = 0 & : \delta_x \\ 1 - 2x'(0) = 0 & : \delta_{x(0)} \\ \cancel{2x'(1) - 2 - \mu = 0} & : \delta_{x(1)} \\ x(1) = 2 & : \delta_\mu \\ \int_0^1 x\, dt = 0 & : \delta_\lambda \end{cases} \quad \text{trivially solved}$$

Part B, question 5

Given the problem

$$\begin{aligned} &\text{minimize:} && \mathcal{F}(x,y,z) = z(0) + \int_0^1 \Big(x^2 + xz + z'^2 + x'y'\Big)\, dt \\ &\text{subject to:} && z(1) = 2 \qquad y(0) = 1 \end{aligned}$$

the differential equations managing the boundary value problem are based on the integral part of $\mathcal{F}$, so on $L(x,y,z) = x^2 + xz + z'^2 + x'y'$. In particular we have

$$\begin{aligned} \frac{\partial L}{\partial x} - \frac{d}{dt}\frac{\partial L}{\partial x'} &= 2x - z - \frac{d}{dt}(y') = && 2x - z - y'' = 0 \\ \frac{\partial L}{\partial y} - \frac{d}{dt}\frac{\partial L}{\partial y'} &= 0 - \frac{d}{dt}(x') = && x'' = 0 \\ \frac{\partial L}{\partial z} - \frac{d}{dt}\frac{\partial L}{\partial z'} &= x - \frac{d}{dt}(2z') = && x - 2z'' = 0 \end{aligned}$$

To determine the other boundary condition is necessary to compute the first variation of the lagrangian $\mathcal{L}(x,y,z,\mu_1,\mu_2) = \mathcal{F} - \mu_1(z(1) - 2) - \mu_2(y(0) - 1)$ that's

$$\begin{aligned}\delta\mathcal{L} = &\int_0^1 \ldots\, dt + \left[\frac{\partial L}{\partial x'}\delta_x\right]_0^1 + \left[\frac{\partial L}{\partial y'}\delta_y\right]_0^1 + \left[\frac{\partial L}{\partial z'}\delta_z\right]_0^1 \\ &+ \delta_{z(0)} - \mu_1\delta_{z(1)} - \big(z(1) - 2\big)\delta_{\mu_1} - \mu_2\delta_{y(0)} - \big(y(0) - 1\big)\delta_{\mu_2} \\ = &\int_0^1 \ldots\, dt + y'(1)\delta_{x(1)} - y'(0)\delta_{x(0)} + x'(1)\delta_{y(1)} - x'(0)\delta_{y(0)} + 2z'(1)\delta_{z(1)} - 2z'(0)\delta_{z(0)} \\ &+ \delta_{z(0)} - \mu_1\delta_{z(1)} - \big(z(1) - 2\big)\delta_{\mu_1} - \mu_2\delta_{y(0)} - \big(y(0) - 1\big)\delta_{\mu_2}\end{aligned}$$

The boundary condition of the boundary value problem can be so obtained by setting equal to zero the terms related to each variation at the extremas, so:

$$\begin{aligned} \delta_{x(0)}: &\quad y'(0) = 0 & \qquad \delta_{x(1)}: &\quad y'(1) = 0 \\ \delta_{y(0)}: &\quad \cancel{x'(0) + \mu_2 = 0} & \qquad \delta_{y(1)}: &\quad x'(1) = 0 \\ \delta_{z(0)}: &\quad 1 - 2z'(0) = 0 & \qquad \delta_{z(1)}: &\quad \cancel{2z'(1) - \mu_2 = 0} \\ \delta_{\mu_1}: &\quad z(1) - 2 = 0 & \qquad \delta_{\mu_2}: &\quad y(0) - 1 = 0 \end{aligned}$$

Part B, question 7

Given the optimal control problem

$$\begin{aligned} \text{minimize:} \quad & x(0) + \int_0^1 xu\,dt \\ \text{subject to:} \quad & x' = y - u \qquad y' = xu \\ & x(1) = 2 \\ & u(t) \in [-2, 1] \end{aligned}$$

the associated hamiltonian $\mathcal{H}$ and utility function $\mathcal{B}$ are

$$\begin{aligned} \mathcal{H}(x, y, u, \lambda_1, \lambda_2) &= xu + \lambda_1(y - u) + \lambda_2 xu \\ \mathcal{B}(x_0, x_1, \mu) &= x_0 + \mu(x_1 - 2) \end{aligned}$$

The co-equation are computed directly from the hamiltonian as

$$\begin{aligned} \lambda_1' &= -\frac{\partial \mathcal{H}}{\partial x} = -(u + \lambda_2 u) \\ \lambda_2' &= -\frac{\partial \mathcal{H}}{\partial y} = -\lambda_1 \end{aligned}$$

The control law can be obtained by the Pontryagin minimum principle, so determining the controls u in it's domain that minimize the function $\tilde{H} = xu - \lambda_1 u + \lambda_2 xu$:

$$u(t) = \underset{\overline{u} \in [-2,1]}{\operatorname{argmin}} \left\{ \overline{u}(x - \lambda_1 + \lambda_2 x) \right\} = \begin{cases} -2 & \text{if } x - \lambda_1 + \lambda_2 x > 0 \\ 1 & \text{if } x - \lambda_1 + \lambda_2 x < 0 \end{cases}$$

The adjoint boundary conditions are instead dependent on the utility function $\mathcal{B}$, in particular

$$\lambda_1(0) = -\frac{\partial \mathcal{B}}{\partial x_0} = -1 \qquad \lambda_2(0) = -\frac{\partial \mathcal{B}}{\partial y_0} = 0 \qquad \lambda_2(1) = \frac{\partial \mathcal{B}}{\partial y_1} = 0$$

Part B, question 8

Given the optimal control problem

$$\begin{aligned} \text{minimize:} \quad & x(1) + 2y(0) + \int_0^1 (1 + x + y)u^2\,dt \\ \text{subject to:} \quad & x' = y \qquad y' = x + u \\ & \int_0^1 (x + u^2)\,dt = 2 \\ & x(0) = 1 \qquad x(1) = 2 \qquad y(0) = 2 \end{aligned}$$

the first thing is to get rid of the integral constraint by adding a new state z whose differential equation is $z' = x + u^2$ and bounded with $z(0) = 0$ and $z(1) = 2$:

$$\begin{aligned} \text{minimize:} \quad & x(1) + 2y(0) + \int_0^1 (1 + x + y)u^2\,dt \\ \text{subject to:} \quad & x' = y \qquad y' = x + u \qquad z' = x + u^2 \\ & x(0) = 1 \qquad x(1) = 2 \qquad y(0) = 2 \qquad z(0) = 0 \qquad z(1) = 2 \end{aligned}$$

With that problem as here presented, the hamiltonian $\mathcal{H}$ and the utility function $\mathcal{B}$ are so

$$\begin{aligned} \mathcal{H}(x, y, z, u, \lambda_1, \lambda_2, \lambda_3) &= (1 + x + y)u^2 + \lambda_1 y + \lambda_2(x + u) + \lambda_3(x + u^2) \\ \mathcal{B}(x_0, x_1, y_0, z_0, z_1, \mu_1, \mu_2, \mu_3, \mu_4, \mu_5) &= x_1 + 2y_0 + \mu_1(x_0 - 1) + \mu_2(x_1 - 2) + \mu_3(y_0 - 2) \\ &\quad + \mu_4 z_0 + \mu_5(z_1 - 2) \end{aligned}$$

Using the Pontryagin maximum principle, we know that the hamiltonian must be stationary respect to the chosen control so

$$\frac{\partial \mathcal{H}}{\partial u} = 2u(1 + x + y) + \lambda_2 + 2u\lambda_3 = 0$$

and inverting to explicit the control

$$u(t) = -\frac{\lambda_2(t)}{2\big(1 + x(t) + y(t) + \lambda_3(t)\big)}$$

The adjoint equations are

$$\begin{aligned}
\lambda_1' &= -\frac{\partial \mathcal{H}}{\partial x} = -\big(u^2 + \lambda_2 + \lambda_3\big) \\
\lambda_2' &= -\frac{\partial \mathcal{H}}{\partial y} = -\big(u^2 + \lambda_1\big) \\
\lambda_3' &= -\frac{\partial \mathcal{H}}{\partial z} = 0
\end{aligned}$$

with adjoint boundary conditions

$$\begin{aligned}
\cancel{\lambda_1(0) = -\frac{\partial \mathcal{B}}{\partial x_0} = -\mu_1} \qquad & \cancel{\lambda_1(1) = \frac{\partial \mathcal{B}}{\partial x_1} = 1 + \mu_2} \\
\cancel{\lambda_2(0) = -\frac{\partial \mathcal{B}}{\partial y_0} = -2 - \mu_3} \qquad & \lambda_2(1) = \frac{\partial \mathcal{B}}{\partial y_1} = 0 \\
\cancel{\lambda_3(0) = -\frac{\partial \mathcal{B}}{\partial z_0} = -\mu_4} \qquad & \cancel{\lambda_3(1) = \frac{\partial \mathcal{B}}{\partial z_1} = \mu_5}
\end{aligned}$$