Chapter 1

Ordinary Differential Equations

In mathematics, an *ordinary differential equation* is a *differential equation* containing one or more functions of one independent variable and its derivatives. The term *ordinary* is used in contrast with the term *partial* so to specify that the differential equations has only *one* independent variable. This mathematical tool is widely used to describe dynamical systems (physical, social, economical, biological and so on and so forth). The first mathematicians to study and apply ordinary differential equations include important names like Newton, Euler, Leibniz, the Bernoulli family, Riccati, Clairaut and d'Alembert. An example of this kind of equations is probably one of the most known formulas in the world, that is trivially the Newton's second law of motion:

$$\mathcal{F}(z(t)) = m\frac{d^2}{dt^2}z(t) \tag{1.1}$$

where $\mathcal{F}$ is a function of the unknown position function $z(t)$ at time t.

1.1 An introduction to Initial Value Problems

Before getting into any numerical solution scheme let us briefly recall the essential formal definitions and properties of ordinary differential equations and initial value problems.

Definition 1 (Ordinary differential equation)**.** An *ordinary differential equation* (ODE) is an equation for a function $z(t)$, defined on an interval $I \subset \mathbb{R}$ and with values in the real or complex numbers or in the space $\mathbb{R}^d$ (or $\mathbb{C}^d$), of the form:

$$F\left(t, z(t), z'(t), z''(t) \ldots, z^{(n)}\right) = 0 \tag{1.2}$$

Here F represents an arbitrary function of its arguments. The *order* n of a differential equation is the highest derivative which occurs. If the dimension d of the value range of $z(t)$ is higher than one, we talk about *systems of differential equations.*

Definition 2. An *explicit* differential equation of first order is a equation of the form:

$$z'(t) = f\left(t, z(t)\right) \qquad \text{or shortly:} \qquad z' = f\left(t, z\right) \tag{1.3}$$

A differential equation of order n is called explicit, if it is of the form:

$$z^{(n)}(t) = F\left(t, z(t), z'(t), z''(t) \ldots, z^{(n-1)}(t)\right) \tag{1.4}$$

Definition 3. A differential equation of the form (1.3) is called *autonomous*, if the right hand side f is not explicitly dependent on t, namely:

$$z'(t) = f\left(z(t)\right) \tag{1.5}$$

Lemma 0.1. *Every differential equation of higher order can be written as a system of first-order differential equations. If the equation is explicit, then the system is explicit.*

Proof. By the introduction of additional variables $z_0(t) = z(t)$, $z_1(t) = z'(t)$ to $z_{n-1}(t) = z^{(n-1)}(t)$, each differential equation of order n can be transformed into a system of n differential equations of first order. This system has the form:

$$\begin{Bmatrix} z_0'(t) - z_1(t) \\ z_1'(t) - z_2(t) \\ \vdots \\ z_{n-2}'(t) - z^{(n-2)}(t) \\ F\left(t, z_0(t), z_1(t), \ldots, z_{n-1}(t), z_{n-1}'(t)\right) \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{Bmatrix} \tag{1.6}$$

In the case of an explicit equation, the system has the form:

$$\begin{Bmatrix} z_0'(t) \\ z_1'(t) \\ \vdots \\ z_{n-2}'(t) \\ z_{n-1}'(t) \end{Bmatrix} = \begin{Bmatrix} z_1(t) \\ z_2(t) \\ \vdots \\ z_{n-1}(t) \\ F\left(t, z_0(t), z_1(t), \ldots, z_{n-1}(t)\right) \end{Bmatrix} \tag{1.7}$$

□

We will see that most of times the ODEs do not come alone. Almost always they are coupled with an *initial condition*, thus forming the so-called *initial value problem* (also known as *Cauchy problem*). The initial condition give information about the value of function z at a given point t_0 in the domain.

Definition 4 (Initial value problem)**.** Given a point $(t_0, u_0) \in \mathbb{R} \times \mathbb{R}^d$. Furthermore, let the function $f(t, z(t))$ with values in $\mathbb{R}^d$ be defined in a neighborhood $I \times U \subset \mathbb{R} \times \mathbb{R}^d$ of the initial value. Then an *initial value problem* (IVP) is defined as follows: find a function $z(t)$, such that:

$$\begin{cases} z'(t) = f(t, z(t)) \\ z(t_0) = z_0 \end{cases} \tag{1.8}$$

Definition 5 (Local solution)**.** We call a continuously differentiable function $z(t)$ with $z(t_0) = z_0$ a *local solution* of the IVP, if there exists a neighborhood J of the point in time t_0 in which $z(t)$ and $f(t, z(t))$ are defined and if the equation $z(t_0) = z_0$ holds for all $t \in J$.

Definition 6 (Linear ODE)**.** A differential equation is said to be *linear* if F can be written as a linear combination of the derivatives of $z(t)$:

$$z^{(n)}(t) = \sum_{i=1}^{n-1} a_i(t) z^{(i)}(t) + r(t) \tag{1.9}$$

with $a_i(t)$ and $r(t)$ continuous functions in t. If $r(t) = 0$ the we call the linear differential equation *homogeneous* otherwise we call it *inhomogeneous*.

1.1.1 Well-posedness of the Initial Value Problems

Definition 7. A mathematical problem is called well-posed if the following *Hadamard conditions* are satisfied:

- A solution exists.
- The solution is unique.
- The solution is continuously dependent on the data.

In the specific case of IVPs, the third condition is often dropped and substituted with the so-called Lipschitz continuity, which is more a quantitative condition.

Definition 8. The function $f(t,z)$ satisfies on its domain $D = I \times \Omega \subset \mathbb{R} \times \mathbb{R}^d$ an uniformly continuous Lipschitz condition if it is Lipschitz continuous with regard to z. In other words it exists a positive constant L, such that:

$$\forall t \in I; x, z \in \Omega : |f(t,x) - f(t,z)| \leq L|x - z| \tag{1.10}$$

It satisfies a local Lipschitz condition if the same holds true for all compact subsets of D.

As we will see the *Peano's existence theorem* proofs that if f is a continuous map than there exists at least a solution to the ODE. It must be pointed out that the Peano theorem does not proof that the solution is unique.

Theorem 1 (Peano's existence theorem)**.** *Let the function $f(t,z)$ be continuous on the closed set*

$$\overline{D} = \left\{(t,z) \in \mathbb{R} \times \mathbb{R}^d \;\middle|\; |t - t_0| \leq \alpha, |z - z_0| \leq \beta\right\} \tag{1.11}$$

where $\alpha, \beta > 0$. Then there exists a solution $z \in \mathbb{C}^1(I)$ on the interval $I = [t_0 - T, t_0 + T]$ with:

$$T = \min\left(\alpha, \frac{\beta}{M}\right), M = \max_{(t,u)\in\overline{D}} |f(t,z)| \tag{1.12}$$

Theorem 2 (Peano's continuation theorem)**.** *Let the assumptions of Peano's existence theorem hold. Then, the solution can be extended to an interval $I_m = [t_-, t_+]$ such that the points $(t_-, z(t_-))$ and $(t_+, z(t_+))$ are on the boundary of D. Neither the values of t, nor of $z(t)$ need to be bounded as long as f remains bounded.*

To proof the uniqueness of solution we employ the aforementioned *Lipschitz continuity* together with the *Grönwall's lemma*.

Lemma 2.2 (Grönwall's theorem)**.** *Let be $w(t)$, $a(t)$ and $b(t)$ be nonnegative, integrable functions, such that $a(t)w(t)$ is integrable. Furthermore, let $b(t)$ be monotonically nondecreasing and let $w(t)$ satisfy the integral inequality:*

$$w(t) \leq b(t) + \int_{t_0}^{t} a(s)w(s)ds, \qquad t \geq t_0 \tag{1.13}$$

Then, for almost all $t \geq t_0$ there holds:

$$w(t) \leq b(t) \exp\left(\int_{t_0}^{t} a(s)ds\right) \tag{1.14}$$

The existence and uniqueness of solution can be now proofed thanks to the *Picard-Lindelöf theorem* (also called *Picard's existence theorem*, *Cauchy-Lipschitz theorem*, or *existence and uniqueness theorem*):

Theorem 3 (Picard-Lindelöf theorem)**.** *Let $f(t,z)$ be continuous on a cylinder $D = (t,z) \in \mathbb{R} \times \mathbb{R}^d \;\middle|\; |t - t_0| \leq a, |z - z_0| \leq b$. Let f be bounded such that there is a constant $M = \max_D |f|$ and satisfy the Lipschitz condition with constant L on D. Then the IVP:*

$$\begin{cases} z'(t) = f(t, z(t)) \\ z(t_0) = z_0 \end{cases} \tag{1.15}$$

is uniquely solvable on the interval $I = [t_0 - T, t_0 + T]$ where $T = \min\{a, b/M\}$.

1.2 Numerical Solutions of Initial Value Problems

In most cases an analytical solution to IVPs cannot be found or it is simply impractical (complex integrals appear in the solution). Thus, a set of numerical schemes for solving IVPs are now presented. It must be pointed out that all the schemes can be naturally extended to systems of differential equations. Moreover, since higher order differential equations can be rewritten as a system of first order differential equations, we will only concentrate on numerical methods for these last ones.

Definition 9 (Time step)**.** On a time interval $I = [t_0, t_0 + T]$, we define a partitioning in n subintervals, also known as *time steps*. The time steps $I_k = [t_{k-1}, t_k]$ have the step size $h_k = t_k - t_{k-1}$. A partitioning in n time steps implies $t_n = T$. The term k-th time step is used for both the interval I_k and for the point in time t_k, but it should always be clear through context which one is meant. Very often, we will consider evenly spaced time steps, in which case we denote the step size by h and $h_k = h$ for all k.

Numerical methods can be subdivided into two main categories: *one-step methods* and *multi-step methods*. In the first category the step z_{k+1} is determined only by z_k, where as in the second category the the step z_{k+1} is determined also by $z_{k-1}, \ldots, z_{k-n}$

Definition 10 (Explicit method)**.** An *explicit method* (forward integration) is a method which, given z_0 at t_0 computes a sequence of approximations $z_1, \ldots, z_n$ to the solution of an IVP in the time steps $t_1, \ldots, t_n$ using an update formula of the form:

$$z_k = z_{k-1} + h_k F(t_{k-1}, z_{k-1}, \ldots, z_{k-p}) \tag{1.16}$$

where F is called *increment function* and p is the order is the numerical multi-step method. If $p = 1$ we call the method as one-step method.

Definition 11 (Implicit method)**.** An *implicit method* (backward integration) is a method which, given z_0 at t_0 computes a sequence of approximations $z_1, \ldots, z_n$ to the solution of an IVP in the time steps $t_1, \ldots, t_n$ using an update formula of the form:

$$z_k = z_{k-1} + h_k F(t_k, z_k, \ldots, z_{k-p}) \tag{1.17}$$

Notice that the increment function F depends on z_k and the previous equation must be solved for z_k.

1.2.1 Euler Method

The first and simplest on-step numerical method is the Euler method. There are many ways of deriving this method. To derive this method let us consider the following IVP:

$$\begin{cases} z'(t) = f(t, z) \\ z(t_0) = z_0 \end{cases} \tag{1.18}$$

The truncated Taylor series of the solution $z(t)$ centered in t_0, that is:

$$z(t_0 + h) = z_0 + hz_1 + o(h^2) \tag{1.19}$$

where $z_1(t = t_0) = z'(t = t_0)$ and since $z(t)$ is considered to be the solution is the IVPs $z_1(t = t_0) = f(t_0, z_0)$. We get an approximation of $z(t_0 + h)$, namely:

(1.20)

Notice that the difference between $\tilde{z}(t_0 + h)$ and $z(t_0 + h)$ is proportional to h^2.

If we consider a generic interval $I = [t_0, t_0 + T]$, we can generalize the equation (1.20) to the k-th step and repeat it on each k-th subinterval $I_k = [t_{k-1}, t_k]$. Thus, in general the k-th step of the *explicit Euler method* can be written as:

$$z(t_k) = z_{k-1} + hf(t_{k-1}, z_{k-1}), \quad k = 1, \ldots, n \tag{1.21}$$

whereas the the k-th step of the *implicit Euler method* can be written as:

$$z(t_k) = z_{k-1} + hf(t_k, z_k), \quad k = 1, \ldots, T/h \tag{1.22}$$

It can be prrofed that both explicit Euler method and implicit Euler method converges to the exact solutionas $h \to 0$ and the error at any time $t \in I = [t_0, t_0 + T]$ can be bounded by Ch, where C is a positive constant. Since C is proportional to e^{LT}, where the Lipschitz constant L of f may be very large. Such problem, which are commonly referred to be *stiff*, are characterized by a large changes in at very different time scales and high sensitivity to changes in the initial condition. If we apply the Euler method to a stiff problems it turns out that the explicit method is not able to properly find the solution due to its *stability*. On the other hand implicit Euler method works large number of applications with a rate of converge of $o(h)$.

1.2.2 Runge-Kutta Methods

Even if the Euler method works fine for most of applications, if the dimension d is large, the end time T is large, the error tolerance ε is small or more importantly the ODE is stiff we might want to improve the solution method. To pursue these improvements we can use the so-called *Runge-Kutta methods*.

Explicit Runge-Kutta Methods

The generic *explicit Runge-Kutta method* is a one-step method with the representation:

$$\begin{aligned} g_i &= z_k + h \sum_{j=1}^{i-1} a_{ij} k_j \\ k_i &= f(hc_i, g_i) \\ g_{k+1} &= z_k + h \sum_{i=1}^{s} b_i k_i \end{aligned} \tag{1.23}$$

where $i = 1, \ldots, s$. In this method the values hc_i are the quadrature points on the interval $[0, h]$. The values k_i are approximations to function values of the integrand in these points and the values gi constitute approximations to the solution $z(hc_i)$ in the quadrature points. This method uses s intermediate values and is thus called an s-stage method.

Definition 12 (Butcher tableau)**.** It is customary to write Runge-Kutta methods in the form of a *Butcher tableau*, containing only the coefficients of equation 1.23 in the following matrix form:

$$\begin{array}{c|cccccc} 0 & 0 & 0 & 0 & 0 & 0 \\ c_2 & a_{21} & 0 & 0 & 0 & 0 \\ c_3 & a_{31} & a_{32} & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 & 0 \\ c_s & a_{s1} & a_{s2} & \ldots & a_{s,s-1} & 0 \\ \hline & b_1 & b_2 & \ldots & b_{s-1} & b_s \end{array} \tag{1.24}$$

Explicit Euler A more generalized formulation of the explicit Euler method.

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array} \tag{1.25}$$

Midpoint method The midpoint method is a variation of the explicit Euler method, also known as Collatz method. It is a second-order method with two stages.

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1/2 & 1/2 & 0 \\ \hline & 0 & 1 \end{array} \tag{1.26}$$

Heun's method The Heun's method is also known as the explicit trapezoid rule since it improves the standard explicit Euler method. It is a second-order method with two stages.

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & 1/2 & 1/2 \end{array} \tag{1.27}$$

Ralston's method It is a second-order method with two stages and a minimum local error bound.

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 2/3 & 2/3 & 0 \\ \hline & 1/4 & 3/4 \end{array} \tag{1.28}$$

Generic second-order method

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \alpha & \alpha & 0 \\ \hline & 1 - \dfrac{1}{2\alpha} & \dfrac{1}{2\alpha} \end{array} \tag{1.29}$$

Runge-Kutta's third-order method (RK3)

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 1 & -1 & 2 & 0 \\ \hline & 1/6 & 2/3 & 1/6 \end{array} \tag{1.30}$$

Heun's third-order method

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 \\ 1/3 & 0 & 2/3 & 0 \\ \hline & 1/4 & 0 & 3/4 \end{array} \tag{1.31}$$

Ralston's third-order method

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 3/4 & 0 & 3/4 & 0 \\ \hline & 2/9 & 1/3 & 4/9 \end{array} \tag{1.32}$$

Strong Stability Preserving Runge-Kutta (SSPRK3)

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1/2 & 1/4 & 1/4 & 0 \\ \hline & 1/6 & 1/6 & 2/3 \end{array} \tag{1.33}$$

Generic third-order method With $\alpha \neq 0, 2/3, 1$:

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \alpha & \alpha & 0 & 0 \\ 1 & 1+\dfrac{1-\alpha}{\alpha(3\alpha-2)} & -\dfrac{1-\alpha}{\alpha(3\alpha-2)} & 0 \\ \hline & \dfrac{1}{2}-\dfrac{1}{6\alpha} & \dfrac{1}{6\alpha(1-\alpha)} & \dfrac{2-3\alpha}{6\alpha(1-\alpha)} \end{array} \tag{1.34}$$

Classical Runge-Kutta's method (RK4) The classical fourth-order Runge-Kutta:

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & 1/6 & 2/6 & 2/6 & 1/6 \end{array} \tag{1.35}$$

Ralston's fourth-order method It is a second-order method with two stages and a minimum truncation error.

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ .4 & .4 & 0 & 0 & 0 \\ .45573725 & .29697761 & .15875964 & 0 & 0 \\ 1 & .21810040 & -3.05096516 & 3.83286476 & 0 \\ \hline & .17476028 & -.55148066 & 1.20553560 & .17118478 \end{array} \tag{1.36}$$

3/8-rule fourth-order method It is a second-order method with two stages and a minimum truncation error.

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 0 \\ 2/3 & -1/3 & 1 & 0 & 0 \\ 1 & 1 & -1 & 1 & 0 \\ \hline & 1/8 & 3/8 & 3/8 & 1/8 \end{array} \tag{1.37}$$

Implicit Runge-Kutta Methods

All the presented method are *explicit Runge-Kutta methods*. If we deal with *stiff* equations the *implicit Runge-Kutta method* should be preferred.

The generic *implicit Runge-Kutta method* is a one-step method with the representation:

$$\begin{aligned} g_i &= z_k + h \sum_{j=1}^{s} a_{ij} k_j \\ k_i &= f(hc_i, g_i) \\ g_{k+1} &= z_k + h \sum_{i=1}^{s} b_i k_i \end{aligned} \tag{1.38}$$

The butcher tableau for (1.38) will be then of the form:

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \dots & a_{1s} \\ c_2 & a_{21} & a_{22} & \dots & a_{2s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & a_{s2} & \dots & a_{ss} \\ \hline & b_1 & b_2 & \dots & b_s s \end{array} \tag{1.39}$$

Implicit Euler method A more generalized formulation of the implicit Euler method.

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array} \tag{1.40}$$

Implicit midpoint method The implicit midpoint method is a variation of the explicit Euler method, also known as Collatz method. It is a second-order method with two stages.

$$\begin{array}{c|c} 1/2 & 1/2 \\ \hline & 1 \end{array} \tag{1.41}$$

Crank-Nicolson method The Crank-Nicolson method corresponds to the implicit trapezoidal rule.

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array} \tag{1.42}$$

Gauss-Legendre methods These methods are based on the points of Gauss-Legendre quadrature.

Gauss-Legendre second-order method

$$\begin{array}{c|cc} \frac{1}{2}-\frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4}-\frac{\sqrt{3}}{6} \\ \frac{1}{2}+\frac{\sqrt{3}}{6} & \frac{1}{4}+\frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \\ & \frac{1}{2}+\frac{\sqrt{3}}{2} & \frac{1}{2}-\frac{\sqrt{3}}{2} \end{array} \tag{1.43}$$

Gauss-Legendre sixth-order method

$$\begin{array}{c|ccc} \frac{1}{2}-\frac{\sqrt{15}}{10} & \frac{5}{36} & \frac{2}{9}-\frac{\sqrt{15}}{15} & \frac{5}{36}-\frac{\sqrt{15}}{30} \\ \frac{1}{2} & \frac{5}{36}+\frac{\sqrt{15}}{24} & \frac{2}{9} & \frac{5}{36}-\frac{\sqrt{15}}{24} \\ \frac{1}{2}+\frac{\sqrt{15}}{10} & \frac{5}{36}+\frac{\sqrt{15}}{30} & \frac{2}{9}+\frac{\sqrt{15}}{15} & \frac{5}{36} \\ \hline & \frac{5}{18} & \frac{4}{9} & \frac{5}{18} \\ & -\frac{5}{6} & \frac{8}{3} & -\frac{5}{6} \end{array} \tag{1.44}$$

Lobatto methods

Lobatto IIIA fourth-order method

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1/2 & 5/24 & 1/3 & -1/24 \\ 1 & 1/6 & 2/3 & 1/6 \\ \hline & 1/6 & 2/3 & 1/6 \end{array} \tag{1.45}$$

Lobatto IIIB fourth-order method

$$\begin{array}{c|ccc} 0 & 1/6 & -1/6 & 0 \\ 1/2 & 1/6 & 1/3 & 0 \\ 1 & 1/6 & 5/6 & 0 \\ \hline & 1/6 & 2/3 & 1/6 \end{array} \tag{1.46}$$

Lobatto IIIC fourth-order method

$$\begin{array}{c|ccc} 0 & 1/6 & -1/3 & 1/6 \\ 1/2 & 1/6 & 5/12 & -1/12 \\ 1 & 1/6 & 2/3 & 1/6 \\ \hline & 1/6 & 2/3 & 1/6 \end{array} \tag{1.47}$$

Lobatto IIIC* fourth-order method

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1/2 & 1/4 & 1/4 & 0 \\ 1 & 0 & 1 & 0 \\ \hline & 1/6 & 2/3 & 1/6 \end{array} \tag{1.48}$$

Radau methods

Radau IA fifth-order method

$$\begin{array}{c|ccc} 0 & \frac{1}{9} & \frac{-1-\sqrt{6}}{18} & \frac{-1+\sqrt{6}}{18} \\ \frac{3}{5}-\frac{\sqrt{6}}{10} & \frac{1}{9} & \frac{11}{45}+\frac{7\sqrt{6}}{360} & \frac{11}{45}-\frac{43\sqrt{6}}{360} \\ \frac{3}{5}+\frac{\sqrt{6}}{10} & \frac{1}{9} & \frac{11}{45}+\frac{43\sqrt{6}}{360} & \frac{11}{45}-\frac{7\sqrt{6}}{360} \\ \hline & \frac{1}{9} & \frac{4}{9}+\frac{\sqrt{6}}{36} & \frac{4}{9}-\frac{\sqrt{6}}{36} \end{array} \tag{1.49}$$

Radau II fifth-order method

$$\begin{array}{c|ccc} \frac{2}{5}-\frac{\sqrt{6}}{10} & \frac{11}{45}-\frac{7\sqrt{6}}{360} & \frac{37}{225}-\frac{169\sqrt{6}}{1800} & -\frac{2}{225}+\frac{\sqrt{6}}{75} \\ \frac{2}{5}+\frac{\sqrt{6}}{10} & \frac{37}{225}+\frac{169\sqrt{6}}{1800} & \frac{11}{45}+\frac{7\sqrt{6}}{360} & -\frac{2}{225}-\frac{\sqrt{6}}{75} \\ 1 & \frac{4}{9}-\frac{\sqrt{6}}{36} & \frac{4}{9}+\frac{\sqrt{6}}{36} & \frac{1}{9} \\ \hline & \frac{4}{9}-\frac{\sqrt{6}}{36} & \frac{4}{9}+\frac{\sqrt{6}}{36} & \frac{1}{9} \end{array} \tag{1.50}$$

Chapter 2

Linear Differential Algebraic Equations

2.1 Linear Differential Algebraic Equations

In this chapter linear implicit differential equations or DAE s of the form

$$\boldsymbol{E}\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{f}(t) \tag{2.1}$$

are studied. In the simplest case, $\boldsymbol{E}, \boldsymbol{A} \in \mathbb{R}^{n\times n}$ are square constant matrices and $\boldsymbol{f}(t) : \mathbb{R} \to \mathbb{R}^n$ is some external perturbation independent of $\boldsymbol{x}(t)$. Such equations occur for example by linearization of autonomous non-linear problems with respect to constant (or critical) solutions, here $\boldsymbol{f}(t)$ is a perturbation that pushes away the solution from the nominal trajectory. Other cases are modelling of linear electrical circuits, simple mechanical systems or, in general, linear systems with additional linear algebraic constraints. The first question that we need to answer, regards existence and uniqueness of a solution $\boldsymbol{x}(t) : \mathbb{R} \to \mathbb{R}^n$. Of course the only interesting case is when $\boldsymbol{E}$ is singular (not invertible). Otherwise it is sufficient to invert $\boldsymbol{E}$ and the DAE s can be written as a ODE s as follows,

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{E}^{-1}\boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{E}^{-1}\boldsymbol{f}(t). \tag{2.2}$$

For eq. (3.2) it is sufficient to assume $\boldsymbol{f}(t)$ piecewise continuous [1] and existence and uniqueness are ensured, see for example [4]. Conversely when matrix $\boldsymbol{E}$ is singular, it is not a priori clear if the solution exists or not. A singular matrix $\boldsymbol{E}$ means that ODE s are mixed with linear algebraic constraints. At this point is important consider some regularity conditions for the pair $(\boldsymbol{E}, \boldsymbol{A})$, as shown in the following example.

Example 1. Consider the DAE s with $\boldsymbol{x}(t) = (x_1(t); x_2(t)) \in \mathbb{R}^2$, $\boldsymbol{f}(t) = 0, \forall t \in [0, \infty)$ and matrices $\boldsymbol{E}, \boldsymbol{A}$ given by

$$\boldsymbol{E} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}. \tag{2.3}$$

This DAE s corresponds to the scalar differential equation $\dot{x}_1(t) = x_2(t)$. Fixing an initial condition $\boldsymbol{x}_0 = (0; 0) \in \mathbb{R}^2$ a solution is provided by

$$\boldsymbol{x}(t) = \int_0^t \begin{bmatrix} x_2(\tau) \\ \alpha(\tau) \end{bmatrix} \mathrm{d}\tau, \tag{2.4}$$

where $\alpha(\tau)$ represents an arbitrary function that satisfies the initial condition. For example let us consider $\alpha(t) = \sin(t)$ which is compatible with the initial condition $(\sin(0) = 0)$,

[1] Weaker conditions for $\boldsymbol{f}(t)$ can be used, for example it is sufficient that $\boldsymbol{f}(t)$ is Lebesque measurable.

then the solution $\boldsymbol{x}(t)$ is

$$\boldsymbol{x}(t) = \begin{bmatrix} -\sin(t) \\ -\cos(t) \end{bmatrix}. \tag{2.5}$$

Nevertheless other choices are possible, i.e. $\alpha(t) = t$, then the solution is

$$\boldsymbol{x}(t) = \begin{bmatrix} t^3/6 \\ t^2/2 \end{bmatrix}. \tag{2.6}$$

It is therefore clear that the solution is not unique and that we must assume some regularity assumptions for the pair $(\boldsymbol{E}, \boldsymbol{A})$. ∘

In order to clarify the existence and uniqueness of a solution for eq. (3.1) we need the following definition.

Definition 13. Matrix pair $(\boldsymbol{E}, \boldsymbol{A})$ is said to be a *regular pencil* if there exist a scalar $\lambda_0 \in \mathbb{R}$, such that

$$\det(\boldsymbol{A} - \lambda_0 \boldsymbol{E}) \neq 0. \tag{2.7}$$

Example 2. The pair of matrices $\boldsymbol{E}, \boldsymbol{A} \in \mathbb{R}^{3\times 3}$ defined below is a non regular (singular) pencil.

$$\boldsymbol{E} = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 0 & 0 \\ 0 & -2 & 0 \end{bmatrix}, \quad \boldsymbol{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{2.8}$$

Since computing the determinant of the pencil is identically zero for every value of $\lambda_0 \in \mathbb{R}$.

$$\det(\boldsymbol{A} - \lambda_0 \boldsymbol{E}) = \det\left(\begin{bmatrix} 1-\lambda_0 & 0 & 0 \\ 2\lambda_0 & 1 & 0 \\ 0 & 2\lambda_0 & 0 \end{bmatrix} \right) = 0 \tag{2.9}$$

∘

The key idea is that if pair (E, A) is a regular pencil, then under an appropriate change of coordinates it is possible to simplify the problem. The idea is made more precise by the following theorem.

Theorem 4 (Weirstrass canonical form)**.** *Matrix pair $\boldsymbol{E}, \boldsymbol{A} \in \mathbb{R}^{n\times n}$ is a regular pencil if, and only if, there exist invertible matrices $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ such that*

$$(\boldsymbol{PEQ}, \boldsymbol{PAQ}) = \left(\begin{bmatrix} \boldsymbol{N} & 0 \\ 0 & \boldsymbol{I} \end{bmatrix}, \begin{bmatrix} \boldsymbol{I} & 0 \\ 0 & \boldsymbol{J} \end{bmatrix} \right), \tag{2.10}$$

where $\boldsymbol{N}$ is an upper triangular nilpotent matrix; $\boldsymbol{J}$ is a matrix in Jordan canonical form and $\boldsymbol{I}$ is the identity matrix of suitable dimensions.

Just for the convenience of the reader we recall the definition of nilpotent matrix.

Definition 14. A square matrix $\boldsymbol{N}$ is said to be nilpotent if there exists an integer $\nu \in \mathbb{N}$ such that $\boldsymbol{N}^\nu = 0$.

Remark 1. The smallest ν that satisfies $\boldsymbol{N}^\nu = 0$ is usually called *degree* of $\boldsymbol{N}$. Notice also that any upper triangular matrix with zeros along the main diagonal is nilpotent. ⌟

Example 3. Matrix $\boldsymbol{N} \in \mathbb{R}^{3\times 3}$ as follows is nilpotent with *degree* $\nu = 3$.

$$\boldsymbol{N} = \begin{bmatrix} 0 & 1 & 3 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{bmatrix}, \quad \boldsymbol{N}^2 = \begin{bmatrix} 0 & 0 & 2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \boldsymbol{N}^3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{2.11}$$

∘

The decoupling form in eq. (3.10) is called *Weierstrass canonical form*, for further details see [5]. Moreover the two matrices have the following structure;

$$\boldsymbol{N} = \begin{bmatrix} \boldsymbol{N}_1 & 0 & \dots & 0 \\ 0 & \boldsymbol{N}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \boldsymbol{N}_{n_1} \end{bmatrix}, \quad \boldsymbol{J} = \begin{bmatrix} \boldsymbol{J}_1 & 0 & \dots & 0 \\ 0 & \boldsymbol{J}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \boldsymbol{J}_{n_2}. \end{bmatrix}. \tag{2.12}$$

Here $\boldsymbol{N}_i$ with $i = 1, \dots, n_1$ and $\boldsymbol{J}_i$ with $i = 1, \dots, n_2$ are Jordan blocks. The blocks $\boldsymbol{N}_i$ are associated to zero eigenvalues of matrix $\boldsymbol{E}$, therefore they appear as follows

$$\boldsymbol{N}_i = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & 1 \\ & & & & 0 \end{bmatrix}, \quad i = 1, \dots, n_1, \tag{2.13}$$

and the dimension of each block $\boldsymbol{N}_i$ is equal to the difference between the algebraic and the geometric multiplicity of the corresponding eigenvalue. Conversely blocks $\boldsymbol{J}_i$ with $i = 1, \dots, n_2$ are classical Jordan blocks with the following structure,

$$\boldsymbol{J}_k = \begin{bmatrix} \lambda_k & 1 & & & \\ & \lambda_k & 1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & 1 \\ & & & & \lambda_k \end{bmatrix}, \quad k = 1, \dots, n_2. \tag{2.14}$$

We now sketch a proof for the Weierstrass canonical form. The proof will be constructive and based on a sequential multiplication of non-singular matrices. Here we prove only the sufficiency part, for the necessity see [5].

Proof. Let us consider the regular pencil pair $(\boldsymbol{E}, \boldsymbol{A})$, surely we can write $(\boldsymbol{E}, \boldsymbol{A} + \lambda_0 \boldsymbol{E} - \lambda_0 \boldsymbol{E})$. Now let us define the matrix $\boldsymbol{T}_0 = (\boldsymbol{A} - \lambda_0 \boldsymbol{E})$, which is non singular by regular pencil assumption. Therefore the quantity $\boldsymbol{T}_0^{-1}(\boldsymbol{E}, \boldsymbol{A} + \lambda_0 \boldsymbol{E} - \lambda_0 \boldsymbol{E}) = \left(\boldsymbol{T}_0^{-1}\boldsymbol{E}, \boldsymbol{I} + \lambda_0 \boldsymbol{T}_0^{-1}\boldsymbol{E}\right)$. Now let us call $\boldsymbol{T}_1$ the non singular transformation matrix such that $\boldsymbol{T}_1^{-1}\boldsymbol{T}_0^{-1}\boldsymbol{E}\boldsymbol{T}_1$ is in Jordan canonical form. Then the pair $\boldsymbol{T}_1^{-1}\left(\boldsymbol{T}_0^{-1}\boldsymbol{E}, \boldsymbol{I} + \lambda_0 \boldsymbol{T}_0^{-1}\boldsymbol{E}\right)\boldsymbol{T}_1$ has the following structure,

$$\left(\begin{bmatrix} \tilde{\boldsymbol{N}} & 0 \\ 0 & \tilde{\boldsymbol{J}} \end{bmatrix}, \begin{bmatrix} \boldsymbol{I} + \lambda_0 \tilde{\boldsymbol{N}} & 0 \\ 0 & \boldsymbol{I} + \lambda_0 \tilde{\boldsymbol{J}} \end{bmatrix} \right). \tag{2.15}$$

Here the block $\tilde{\boldsymbol{N}}$ is associated to the zero eigenvalues of the singular matrix $\boldsymbol{T}_0^{-1}\boldsymbol{E}$. Therefore $\tilde{\boldsymbol{N}}$ is upper triangular with zero elements on the main diagonal. Thanks to this

property the matrix $(\boldsymbol{I} + \lambda_0 \tilde{\boldsymbol{N}})$ is instead invertible. Therefore we can consider the non singular transformation matrix $\boldsymbol{T}_2$ defined as follows,

$$\boldsymbol{T}_2 = \begin{bmatrix} (\boldsymbol{I} + \lambda_0 \tilde{\boldsymbol{N}}) & 0 \\ 0 & \tilde{\boldsymbol{J}} \end{bmatrix}. \tag{2.16}$$

Now left multiplying eq. (3.15) by $\boldsymbol{T}_2^{-1}$ we obtain the following structure,

$$\left(\begin{bmatrix} (\boldsymbol{I} + \lambda_0 \tilde{\boldsymbol{N}})^{-1} \tilde{\boldsymbol{N}} & 0 \\ 0 & \boldsymbol{I} \end{bmatrix}, \begin{bmatrix} \boldsymbol{I} & 0 \\ 0 & \tilde{\boldsymbol{J}}^{-1}(\boldsymbol{I} + \lambda_0 \tilde{\boldsymbol{J}}) \end{bmatrix} \right). \tag{2.17}$$

Finally it is sufficient to consider matrices $\boldsymbol{T}_{31}$ and $\boldsymbol{T}_{32}$ such that $\boldsymbol{T}_{31}^{-1}(\boldsymbol{I} + \lambda_0 \tilde{\boldsymbol{N}})^{-1} \tilde{\boldsymbol{N}} \boldsymbol{T}_{31}$ and $\boldsymbol{T}_{32}^{-1} \tilde{\boldsymbol{J}}^{-1}(\boldsymbol{I} + \lambda_0 \tilde{\boldsymbol{J}}) \boldsymbol{T}_{32}$ are in Jordan canonical form and build matrix $\boldsymbol{T}_3$ as follows,

$$\boldsymbol{T}_3 = \begin{bmatrix} \boldsymbol{T}_{31} & 0 \\ 0 & \boldsymbol{T}_{32} \end{bmatrix}. \tag{2.18}$$

Thus the final change of coordinates yields,

$$\boldsymbol{T}_3^{-1} \left(\begin{bmatrix} (\boldsymbol{I} + \lambda_0 \tilde{\boldsymbol{N}})^{-1} \tilde{\boldsymbol{N}} & 0 \\ 0 & \boldsymbol{I} \end{bmatrix}, \begin{bmatrix} \boldsymbol{I} & 0 \\ 0 & \tilde{\boldsymbol{J}}^{-1}(\boldsymbol{I} + \lambda_0 \tilde{\boldsymbol{J}}) \end{bmatrix} \right) \boldsymbol{T}_3 = \left(\begin{bmatrix} \boldsymbol{N} & 0 \\ 0 & \boldsymbol{I} \end{bmatrix}, \begin{bmatrix} \boldsymbol{I} & 0 \\ 0 & \boldsymbol{J} \end{bmatrix} \right). \tag{2.19}$$

It remains to prove that matrix $\boldsymbol{N}$ is nilpotent and to do that is sufficient to show that matrix $(\boldsymbol{I} + \lambda_0 \tilde{\boldsymbol{N}})^{-1} \tilde{\boldsymbol{N}}$ is nilpotent as well. First notice that matrices $(\boldsymbol{I} + \lambda_0 \tilde{\boldsymbol{N}})$ and $\tilde{\boldsymbol{N}}$ commute since,

$$(\boldsymbol{I} + \lambda_0 \tilde{\boldsymbol{N}}) \tilde{\boldsymbol{N}} = \tilde{\boldsymbol{N}} + \lambda_0 \tilde{\boldsymbol{N}}^2 = \tilde{\boldsymbol{N}}(\boldsymbol{I} + \lambda_0 \tilde{\boldsymbol{N}}). \tag{2.20}$$

Then also $(\boldsymbol{I} + \lambda_0 \tilde{\boldsymbol{N}})^{-1}$ and $\tilde{\boldsymbol{N}}$ commute since,

$$(\boldsymbol{I} + \lambda_0 \tilde{\boldsymbol{N}})^{-1} \tilde{\boldsymbol{N}} = (\boldsymbol{I} + \lambda_0 \tilde{\boldsymbol{N}})^{-1} \tilde{\boldsymbol{N}} (\boldsymbol{I} + \lambda_0 \tilde{\boldsymbol{N}})(\boldsymbol{I} + \lambda_0 \tilde{\boldsymbol{N}})^{-1} = \tilde{\boldsymbol{N}}(\boldsymbol{I} + \lambda_0 \tilde{\boldsymbol{N}})^{-1} \tag{2.21}$$

where we used the fact that $\tilde{\boldsymbol{N}}(\boldsymbol{I} + \lambda_0 \tilde{\boldsymbol{N}}) = (\boldsymbol{I} + \lambda_0 \tilde{\boldsymbol{N}}) \tilde{\boldsymbol{N}}$. Now since we know that commutative property for these matrices holds, called ν the index such that $\tilde{\boldsymbol{N}}^\nu = 0$ the following is true

$$\left[(\boldsymbol{I} + \lambda_0 \tilde{\boldsymbol{N}}) \tilde{\boldsymbol{N}} \right]^\nu = (\boldsymbol{I} + \lambda_0 \tilde{\boldsymbol{N}})^\nu \tilde{\boldsymbol{N}}^\nu = 0, \tag{2.22}$$

and this complete the proof. Just to summarize, we notice that the overall transformation matrices $(\boldsymbol{P}, \boldsymbol{Q})$ are given by $\boldsymbol{P} = (\boldsymbol{T}_0 \boldsymbol{T}_1 \boldsymbol{T}_2 \boldsymbol{T}_3)^{-1}$ and $\boldsymbol{Q} = \boldsymbol{T}_1 \boldsymbol{T}_3$. □

The Weirstrass canonical form is quite useful since we can re-write system eq. (3.1) using the transformation matrices $(\boldsymbol{P}, \boldsymbol{Q})$ just introduced, thus

$$\boldsymbol{P}\boldsymbol{E}\boldsymbol{Q}\boldsymbol{Q}^{-1}\dot{\boldsymbol{x}}(t) = \boldsymbol{P}\boldsymbol{A}\boldsymbol{Q}\boldsymbol{Q}^{-1}\boldsymbol{x}(t) + \boldsymbol{P}\boldsymbol{f}(t). \tag{2.23}$$

Now considering the following change of coordinates,

$$\boldsymbol{z}(t) = \boldsymbol{Q}^{-1}\boldsymbol{x}(t), \tag{2.24a}$$
$$\boldsymbol{g}(t) = \boldsymbol{P}\boldsymbol{f}(t), \tag{2.24b}$$

we obtain the following equivalent DAE s

$$\begin{bmatrix} \boldsymbol{N} & 0 \\ 0 & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \dot{\boldsymbol{z}}_1(t) \\ \dot{\boldsymbol{z}}_2(t) \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & 0 \\ 0 & \boldsymbol{J} \end{bmatrix} \begin{bmatrix} \boldsymbol{z}_1(t) \\ \boldsymbol{z}_2(t) \end{bmatrix} + \begin{bmatrix} \boldsymbol{g}_1(t) \\ \boldsymbol{g}_2(t) \end{bmatrix}. \tag{2.25}$$

Here we partitioned the $\boldsymbol{z}(t)$ and $\boldsymbol{g}(t)$ vectors according respectively to the dimensions of $\boldsymbol{J}$ and $\boldsymbol{N}$. Obviously $\boldsymbol{z}(t) = (\boldsymbol{z}_1(t); \boldsymbol{z}_2(t))$ and $\boldsymbol{g}(t) = (\boldsymbol{g}_1(t); \boldsymbol{g}_2(t))$. It is also interesting, notice that the two sets of equations corresponding to $\boldsymbol{z}_1(t)$ and $\boldsymbol{z}_2(t)$ are decoupled

$$\boldsymbol{N}\dot{\boldsymbol{z}}_1 = \boldsymbol{z}_1 + \boldsymbol{g}_1(t), \tag{2.26a}$$

$$\dot{\boldsymbol{z}}_2 = \boldsymbol{J}\boldsymbol{z}_2 + \boldsymbol{g}_2(t). \tag{2.26b}$$

It is straightforward to see that the equation associated to $\boldsymbol{z}_2(t)$ is just an ODE then the unique solution from the initial condition $\boldsymbol{z}_2(0)$ is given by

$$\boldsymbol{z}_2(t) = e^{\boldsymbol{J}t}\boldsymbol{z}_2(0) + \int_0^t e^{\boldsymbol{J}(t-\tau)}\boldsymbol{g}_2(\tau)\,\mathrm{d}\tau, \quad \forall t \geq 0, \tag{2.27}$$

(for further details see for example [6]). eq. (3.26a) is instead again a DAE, but since matrix N is nilpotent we can manipulate it. It is convenient to re-write eq. (3.26a) as

$$\boldsymbol{z}_1(t) = \boldsymbol{N}\dot{\boldsymbol{z}}_1(t) - \boldsymbol{g}_1(t). \tag{2.28}$$

Now substituting eq. (3.28) inside itself and differentiating yields,

$$\begin{aligned}
\boldsymbol{z}_1(t) &= \boldsymbol{N}\frac{\mathrm{d}}{\mathrm{d}t}\left[\boldsymbol{N}\dot{\boldsymbol{z}}_1(t) - \boldsymbol{g}_1(t)\right] - \boldsymbol{g}_1(t) = \boldsymbol{N}^2\frac{\mathrm{d}^2\boldsymbol{z}_1(t)}{\mathrm{d}t^2} - \boldsymbol{N}\frac{\mathrm{d}\boldsymbol{g}_1(t)}{\mathrm{d}t} - \boldsymbol{g}_1(t) \\
&= \boldsymbol{N}^2\frac{\mathrm{d}^2}{\mathrm{d}t^2}\left[\boldsymbol{N}\dot{\boldsymbol{z}}_1(t) - \boldsymbol{g}_1(t)\right] - \boldsymbol{N}\frac{\mathrm{d}\boldsymbol{g}_1(t)}{\mathrm{d}t} - \boldsymbol{g}_1(t) = \boldsymbol{N}^3\frac{\mathrm{d}^3\boldsymbol{z}_1(t)}{\mathrm{d}t^3} - \boldsymbol{N}^2\frac{\mathrm{d}^2\boldsymbol{g}_1(t)}{\mathrm{d}t^2} - \dots \\
&\vdots \\
&= \boldsymbol{N}^\nu\frac{\mathrm{d}^\nu\boldsymbol{z}_1(t)}{\mathrm{d}t^\nu} - \boldsymbol{N}^{\nu-1}\frac{\mathrm{d}^{\nu-1}\boldsymbol{g}_1(t)}{\mathrm{d}t^{\nu-1}} - \dots - \boldsymbol{N}\frac{\mathrm{d}\boldsymbol{g}_1(t)}{\mathrm{d}t} - \boldsymbol{g}_1(t),
\end{aligned}$$

establishing differential properties of the solutions. Notice that the procedure stops when we reach the degree ν of matrix $\boldsymbol{N}$, indeed $\boldsymbol{N}^\nu = 0$ and we may find a unique solution expressed as a finite summation of increasing derivatives of $\boldsymbol{g}(t)$ as follows:

$$\boldsymbol{z}_1(t) = -\sum_{k=0}^{\nu-1}\boldsymbol{N}^k\frac{\mathrm{d}^k\boldsymbol{g}_1(t)}{\mathrm{d}t^k}. \tag{2.30}$$

Remark 2. Sometimes index ν is also called *Kronecker index*. Notice also that ν is the number of times that we differentiated function $\boldsymbol{g}_1(t)$ in order to obtain the solution $\boldsymbol{z}_1(t)$. ⌟

Finally we are ready to express the unique solution to the original problem in eq. (3.1). It is sufficient to invert the change of coordinates in eq. (3.24a) and we obtain

$$\boldsymbol{x}(t) = \boldsymbol{Q}\boldsymbol{z}(t) = \boldsymbol{Q}\begin{bmatrix} -\sum_{k=0}^{\nu-1}\boldsymbol{N}^k\frac{\mathrm{d}^k\boldsymbol{g}_1(t)}{\mathrm{d}t^k} \\ e^{\boldsymbol{J}t}\boldsymbol{z}_2(0) + \int_0^t e^{\boldsymbol{J}(t-\tau)}\boldsymbol{g}_2(\tau)\,\mathrm{d}\tau \end{bmatrix} \tag{2.31}$$

Remark 3. Notice that the initial condition $\boldsymbol{z}_1(0)$ cannot be chosen freely, but has to satisfy the following,

$$\boldsymbol{z}_1(0) = -\sum_{k=0}^{\nu-1}\boldsymbol{N}^k\frac{\mathrm{d}^k\boldsymbol{g}_1(0)}{\mathrm{d}t^k}. \tag{2.32}$$

Therefore also the initial condition $\boldsymbol{x}_0 = \boldsymbol{x}(t=0)$ for eq. (3.1) cannot be chosen completely freely. This intuitively expresses the fact that also the initial condition has to be consistent with constrains expressed by $\boldsymbol{E}$. ⌟

Summarizing what we discussed we have the following,

Theorem 5. *Problem in eq.* (3.1) *with initial condition* $\boldsymbol{x}_0 = \boldsymbol{x}(t=0)$ *has a solution if the following conditions holds:*

- *function* $\boldsymbol{f}(t)$ *is piece-wise continuous,*
- *the pair* $(\boldsymbol{E}, \boldsymbol{A})$ *is a regular pencil,*
- *the initial condition* $\boldsymbol{x}_0$ *satisfies* $\boldsymbol{x}_0 = \boldsymbol{Q}\begin{bmatrix} -\sum_{k=0}^{\nu-1} \boldsymbol{N}^k \frac{\mathrm{d}^k \boldsymbol{g}_1(0)}{\mathrm{d}t^k} \\ \boldsymbol{z}_2(0) \end{bmatrix}$.

Moreover the corresponding solutions is given by eq. (3.31).

2.2 Review of Gaussian Elimination

The Weirstrass canonical form is very useful from the theoretical point of view, however is computationally demanding, because it requires to go through several Jordan canonical forms. A numerical more efficient algorithm to reduce a DAE to an ODE is presented in the next section. The idea is to use the celebrated Gauss's elimination method combined with differentiation in order to perform the desired reduction. Before doing so, we want recall some elementary concepts and definitions useful for the purpose.

Definition 15. Given a matrix $\boldsymbol{M} \in \mathbb{R}^{n\times m}$ we call elementary operations the followings,

- permutation of two rows (resp. columns),
- linear combinations of rows (resp. columns).

These operations can be also regarded as matrix multiplications.

Definition 16. A non-singular matrix $\boldsymbol{P}(i,j) \in \{0,1\}^{n\times n}$ (resp. $\boldsymbol{P}(i,j) \in \{0,1\}^{m\times m}$) is said to be a rows (resp. columns) permutation matrix if it is obtained by the identity matrix $\boldsymbol{I} \in \mathbb{R}^{n\times n}$ (resp. $\boldsymbol{I} \in \mathbb{R}^{m\times m}$) exchanging the column i with column j.

Permutation matrices have the following remarkable properties,

- $\boldsymbol{P}(i,j)^{-1} = \boldsymbol{P}(i,j)^\intercal$, therefore $\boldsymbol{P}(i,j)$ is an orthogonal matrix.
- $\boldsymbol{P}(i,j)^2 = \boldsymbol{I}$, therefore $\boldsymbol{P}(i,j)$ is an involution matrix.
- $\boldsymbol{P}(i,j)\boldsymbol{M}$ exchanges row i with row j.
- $\boldsymbol{M}\boldsymbol{P}(i,j)$ exchanges column i with column j.

Example 4. Consider the following permutation matrix $\boldsymbol{P}(1,2) \in \{1,0\}^{3\times 3}$ and a matrix $\boldsymbol{M} \in \mathbb{R}^{3\times 3}$ then $\boldsymbol{P}(1,2)\boldsymbol{M}$ exchanges rows $(1,2)$ of matrix $\boldsymbol{M}$.

$$\boldsymbol{P}(1,2)\boldsymbol{M} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} = \begin{bmatrix} 4 & 5 & 6 \\ 1 & 2 & 3 \\ 7 & 8 & 9 \end{bmatrix} \tag{2.33}$$

In the same way $\boldsymbol{M}\boldsymbol{P}(1,2)$ exchanges columns $(1,2)$ of matrix $\boldsymbol{M}$,

$$\boldsymbol{M}\boldsymbol{P}(1,2) = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 3 \\ 5 & 4 & 6 \\ 8 & 7 & 9 \end{bmatrix} \tag{2.34}$$

∘

In analogous way it is possible to define a matrix that performs linear combinations of rows (resp. columns). These matrices are a special class of Frobenius matrices and usually are called elementary matrices.

Definition 17. A matrix $\boldsymbol{F}(i,j,\alpha) \in \mathbb{R}^{n\times n}$ (resp. $\boldsymbol{F}(i,j,\alpha) \in \mathbb{R}^{m\times m}$) is said to be a rows (resp. columns) elementary matrix if it has the following structure,

$$\boldsymbol{F}(i,j,\alpha) = \begin{bmatrix} 1 & \dots & 0 & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \dots & 1 & & 0 & & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \dots & \alpha & \dots & 1 & \dots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & 0 & \dots & 1 \end{bmatrix} \tag{2.35}$$

with the (i,j) element equal to α.

Remark 4. An elementary matrix $\boldsymbol{F}(i,j,\alpha) \in \mathbb{R}^{n\times n}$ (resp. $\boldsymbol{F}(i,j,\alpha) \in \mathbb{R}^{m\times m}$) is equal to the identity matrix of dimension n (resp. m) for all the entries except the (i,j) position which is equal to α. ⌟

Combination matrices have the following remarkable properties:

- $\boldsymbol{F}(i,j,\alpha)\boldsymbol{M}$ add the row j multiplied by α to row i,
- $\boldsymbol{M}\boldsymbol{F}(i,j,\alpha)$ add the column i multiplied by α to column j.

Theorem 6. *Given a matrix $\boldsymbol{M} \in \mathbb{R}^{n\times m}$ with $n \leq m$ and* rank $\boldsymbol{M} = r \leq n$*, there exist a pair of non singular matrices, $\boldsymbol{L} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{U} \in \mathbb{R}^{m\times m}$ such that,*

$$\boldsymbol{L}\boldsymbol{M}\boldsymbol{U} = \begin{bmatrix} \boldsymbol{\Lambda} & 0 \\ 0 & 0 \end{bmatrix}, \tag{2.36}$$

with $\boldsymbol{\Lambda} \in \mathbb{R}^{r\times r}$ a non singular diagonal matrix. Blocks of zeros have dimensions compatible with the partitioning.

Proof. The proof is quite algorithmic and based on sequentially Gauss's eliminations. We define $\boldsymbol{M}^{(k)}$ the matrix $\boldsymbol{M}$ at the k iteration. Parenthesis are used to stress out that $\boldsymbol{M}^{(k)}$ is not the matrix $\boldsymbol{M}$ to the power k. So let us initialize the algorithm with $\boldsymbol{M}^{(1)} = \boldsymbol{M}$ and $k = 1$. Now the idea is to look for the firs non zero element inside the minor $\boldsymbol{M}(k \dots n, k \dots m)$; obviously at the first iteration we are exploring all the entries. Suppose to move first along rows and next along columns; then unless the matrix is completely full of zeros we will find a pair of indices such that $m^{(k)}_{pq} \neq 0$. Then the idea is to move this non zero element in the (k,k) position. This can be done with a rows and columns permutations as follows,

$$\overline{\boldsymbol{M}}^{(k)} = \boldsymbol{P}(k,p)\boldsymbol{M}^{(k)}\boldsymbol{P}(q,k). \tag{2.37}$$

Notice that the $\overline{m}^{(k)}_{k,k} = m^{k}_{p,q}$, roughly speaking the non zero element in position (p,q) of matrix $\boldsymbol{M}^{(k)}$ is moved to the (k,k) position of matrix $\overline{\boldsymbol{M}}^{(k)}$. Now we can use this non zero element to set to zero all the element along the k column and the k row. This can be done using a composition of linear combination matrices. Specifically the matrix

$\overline{\boldsymbol{L}}^{(k)} \in \mathbb{R}^{n\times n}$ defined as follows is designed to set to zero all the elements in the indices range $(k\dots n, k\dots k)$,

$$\begin{aligned}\overline{\boldsymbol{L}}^{(k)} &= \boldsymbol{F}\left(k+1,k,-\frac{\overline{m}_{k+1,k}}{\overline{m}_{k,k}}\right)\dots\boldsymbol{F}\left(n-1,k,-\frac{\overline{m}_{n-1,k}}{\overline{m}_{k,k}}\right)\boldsymbol{F}\left(n,k,-\frac{\overline{m}_{n,k}}{\overline{m}_{k,k}}\right)\\ &= \prod_{h=k+1}^{n}\boldsymbol{F}\left(h,k,-\frac{\overline{m}_{h,k}}{\overline{m}_{k,k}}\right)\end{aligned}. \tag{2.38}$$

The same operation can be performed along columns, in this case matrix $\overline{\boldsymbol{U}}^{(k)}$ performs linear combinations to set to zero all the element in the indices range $(k\dots k, k\dots m)$,

$$\begin{aligned}\overline{\boldsymbol{U}}^{(k)} &= \boldsymbol{F}\left(k,k+1,-\frac{\overline{m}_{k,k+1}}{\overline{m}_{k,k}}\right)\dots\boldsymbol{F}\left(k,n-1,-\frac{\overline{m}_{k,n-1}}{\overline{m}_{k,k}}\right)\boldsymbol{F}\left(k,n,-\frac{\overline{m}_{k,n}}{\overline{m}_{k,k}}\right)\\ &= \prod_{h=k+1}^{m}\boldsymbol{F}\left(k,h,-\frac{\overline{m}_{k,h}}{\overline{m}_{k,k}}\right).\end{aligned} \tag{2.39}$$

Now just for notational convenience let us define the following quantities,

$$\boldsymbol{L}^{(k)} = \overline{\boldsymbol{L}}^{(k)}\boldsymbol{P}(k,p), \tag{2.40a}$$

$$\boldsymbol{U}^{(k)} = \boldsymbol{P}(q,k)\overline{\boldsymbol{U}}^{(k)}. \tag{2.40b}$$

Then using simultaneously the two matrices $\boldsymbol{L}^{(k)} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{U}^{(k)} \in \mathbb{R}^{m\times m}$ we can set to zero all the elements along the k row and the k column, obviously except the (k,k) position. Then the matrix M at next iteration is given by,

$$\boldsymbol{M}^{(k+1)} = \boldsymbol{L}^{(k)}\boldsymbol{M}^{(k)}\boldsymbol{U}^{(k)}. \tag{2.41}$$

Is now easy to see that iterating the process; after exactly r iterations, matrix $\boldsymbol{M}^{(r+1)} \in \mathbb{R}^{n\times m}$ has exactly the following structure,

$$\boldsymbol{M}^{(r+1)} = \begin{bmatrix}\boldsymbol{\Lambda} & 0\\ 0 & 0\end{bmatrix}. \tag{2.42}$$

Moreover the matrices $\boldsymbol{L} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{U} \in \mathbb{R}^{m\times m}$ that realize that structure are given by,

$$\boldsymbol{L} = \boldsymbol{L}^{(r)}\boldsymbol{L}^{(r-1)}\dots\boldsymbol{L}^{(1)}, \tag{2.43a}$$

$$\boldsymbol{U} = \boldsymbol{U}^{(1)}\boldsymbol{U}^{(2)}\dots\boldsymbol{U}^{(r)}. \tag{2.43b}$$

Notice also that $\boldsymbol{L}$ and $\boldsymbol{U}$ are non singular since are obtained through product of non singular matrices. □

To make things clear we propose the following example of the algorithm.

Example 5. Let us consider the matrix $\boldsymbol{M} \in \mathbb{R}^{3\times 4}$ with rank $\boldsymbol{M} = r = 2$ defined as follows,

$$\boldsymbol{M} = \begin{bmatrix}0 & 0 & 1 & 0\\ 0 & 2 & 0 & 0\\ 0 & 4 & 2 & 0\end{bmatrix}. \tag{2.44}$$

First of all we set $\boldsymbol{M}^{(1)} = \boldsymbol{M}$ and $k = 1$ and we look for the firs non zero element starting from position $(k,k) = (1,1)$ and moving along the first column and then along the second and so forth. At the first iteration the first non zero element is in the position $(p,q) = (2,2)$,

so the following permutation matrices $\boldsymbol{P}(1,2) \in \{0,1\}^{3\times 3}$ and $\boldsymbol{P}(2,1) \in \{0,1\}^{4\times 4}$ provide a way to move it in the $(1,1)$ position.

$$\overline{\boldsymbol{M}}^{(1)} = \boldsymbol{P}(1,2)\boldsymbol{M}^{(1)}\boldsymbol{P}(2,1) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \boldsymbol{M}^{(1)} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 4 & 0 & 2 & 0 \end{bmatrix} \tag{2.45}$$

Now we need to set to zero all the non zero elements in column and row 1, and the non only non zero element is $\overline{m}^{(1)}_{3,1} = 4$. Therefore the row combination matrix $\boldsymbol{C}(3,1,-4/2) \in \mathbb{R}^{3\times 3}$ reaches the goal,

$$\boldsymbol{M}^{(2)} = \boldsymbol{C}(3,1,-4/2)\overline{\boldsymbol{M}}^{(1)} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -2 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 4 & 0 & 2 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 2 & 0 \end{bmatrix} \tag{2.46}$$

Now $k = 2$ and we perform again a search for the first non zero element in $\boldsymbol{M}^{(2)}$. The answer is $(p,q) = (2,3)$ so we can use again permutation matrices, $\boldsymbol{P}(2,2) = I \in \{0,1\}^{3\times 3}$ and $\boldsymbol{P}(3,1) \in \{0,1\}^{4\times 4}$.

$$\overline{\boldsymbol{M}}^{(2)} = \boldsymbol{M}^{(2)}\boldsymbol{P}(3,1) = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 2 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 \end{bmatrix} \tag{2.47}$$

Finally we can use the rows combination matrix $\boldsymbol{C}(3,2,-2/1) \in \mathbb{R}^{3\times 3}$ to complete the procedure,

$$\boldsymbol{M}^{(3)} = \boldsymbol{C}(3,2,-2/1)\overline{\boldsymbol{M}}^{(2)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -2 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{2.48}$$

Here obviously the diagonal block $\boldsymbol{\Lambda} \in \mathbb{R}^{2\times 2} = \mathrm{diag}(2,1)$. Notice also that is now clear that the rank of matrix M is equal $r = 2$. ∘

2.3 Gauss like method to compute DAEs' index

We are finally ready to discuss why theorem 9 is useful in computing the index of a DAE, and how it can be used to transform a generic linear DAE into an ODE. So let us consider a generic linear DAE in the form shown in eq. (3.1), and we notice that can be also written as,

$$\begin{bmatrix} \boldsymbol{E} & -\boldsymbol{A} \end{bmatrix} \begin{bmatrix} \dot{\boldsymbol{x}} \\ \boldsymbol{x} \end{bmatrix} = \boldsymbol{f}. \tag{2.49}$$

Just for notational brevity we dropped the time dependency. Here matrix $\begin{bmatrix} \boldsymbol{E} & -\boldsymbol{A} \end{bmatrix} \in \mathbb{R}^{n\times 2n}$. Let us call

$$\text{rank } \boldsymbol{E} =: r \leq n. \tag{2.50}$$

Then let us consider a pair of matrices $\boldsymbol{L}^{(0)} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{U}^{(0)} \in \mathbb{R}^{n\times n}$ such that,

$$\boldsymbol{U}^{(0)}\boldsymbol{E}\boldsymbol{L}^{(0)} = \boldsymbol{E}^{(0)} = \begin{bmatrix} \boldsymbol{\Lambda}^{(0)} & 0 \\ 0 & 0 \end{bmatrix}.$$

As done in the previous section we used the superscript $^{(0)}$ to indicate the iteration number; it will be clear soon that this number is also the Kronecker index of the DAE. Notice that at this first iteration $\boldsymbol{\Lambda}^{(0)} \in \mathbb{R}^{r\times r}$. If $r = n$ then we done since the DAE has index $\nu = 0$ and matrix $\boldsymbol{E}$ was full rank and invertible. If $r < n$ it is convenient introduce the following change of coordinates,

$$\begin{bmatrix} \dot{\boldsymbol{x}}^{(0)} \\ \boldsymbol{x}^{(0)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{U}^{(0)} & 0 \\ 0 & \boldsymbol{I} \end{bmatrix}^{-1} \begin{bmatrix} \dot{\boldsymbol{x}} \\ \boldsymbol{x} \end{bmatrix}, \tag{2.51}$$

and also the matrix $\boldsymbol{A}^{(0)}$ and vector $\boldsymbol{f}^{(0)}$ as follows,

$$\begin{aligned} \boldsymbol{A}^{(0)} &:= -\boldsymbol{L}^{(0)}\boldsymbol{A}, \\ \boldsymbol{f}^{(0)} &:= \boldsymbol{L}^{(0)}\boldsymbol{f}. \end{aligned}$$

Let us consider to multiply on the left-hand side eq. (3.49) by $\boldsymbol{L}^{(0)}$. Then we obtain the following equivalent form,

$$\begin{aligned} \boldsymbol{L}^{(0)} \begin{bmatrix} \boldsymbol{E} & -\boldsymbol{A} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}^{(0)} & 0 \\ 0 & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}^{(0)} & 0 \\ 0 & \boldsymbol{I} \end{bmatrix}^{-1} \begin{bmatrix} \dot{\boldsymbol{x}} \\ \boldsymbol{x} \end{bmatrix} &= \begin{bmatrix} \boldsymbol{L}^{(0)}\boldsymbol{E}\boldsymbol{U}^{(0)} & \boldsymbol{L}^{(0)}\boldsymbol{A} \end{bmatrix} \begin{bmatrix} \dot{\boldsymbol{x}}^{(0)} \\ \boldsymbol{x}^{(0)} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{E}^{(0)} & \boldsymbol{A}^{(0)} \end{bmatrix} \begin{bmatrix} \dot{\boldsymbol{x}}^{(0)} \\ \boldsymbol{x}^{(0)} \end{bmatrix} = \boldsymbol{f}^{(0)}(t) \end{aligned}$$

Now notice that matrix $\begin{bmatrix} \boldsymbol{E}^{(0)} & \boldsymbol{A}^{(0)} \end{bmatrix} \in \mathbb{R}^{n\times 2n}$ has the following special structure,

$$\begin{bmatrix} \boldsymbol{\Lambda}^{(0)} & 0 & \boldsymbol{A}_{11}^{(0)} & \boldsymbol{A}_{12}^{(0)} \\ 0 & 0 & \boldsymbol{A}_{21}^{(0)} & \boldsymbol{A}_{22}^{(0)} \end{bmatrix} \begin{bmatrix} \dot{\boldsymbol{x}}_1^{(0)} \\ \dot{\boldsymbol{x}}_2^{(0)} \\ \boldsymbol{x}_1^{(0)} \\ \boldsymbol{x}_2^{(0)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{f}_1^{(0)} \\ \boldsymbol{f}_2^{(0)} \end{bmatrix}. \tag{2.53}$$

Obviously eq. (3.53) can be also written as a couple of matrix equations as follows,

$$\boldsymbol{\Lambda}^{(0)}\dot{\boldsymbol{x}}_1^{(0)} + \boldsymbol{A}_{11}^{(0)}\boldsymbol{x}_1^{(0)} + \boldsymbol{A}_{12}^{(0)}\boldsymbol{x}_2^{(0)} = \boldsymbol{f}_1^{(0)}, \tag{2.54a}$$

$$\boldsymbol{A}_{21}^{(0)}\boldsymbol{x}_1^{(0)} + \boldsymbol{A}_{22}^{(0)}\boldsymbol{x}_2^{(0)} = \boldsymbol{f}_2^{(0)}, \tag{2.54b}$$

where $\dot{\boldsymbol{x}}_1^{(0)}, \boldsymbol{x}_1^{(0)} \in \mathbb{R}^r$, $\dot{\boldsymbol{x}}_2^{(0)}, \boldsymbol{x}_2^{(0)} \in \mathbb{R}^{n-r}$ and $\boldsymbol{x}^{(0)} = [\boldsymbol{x}_1^{(0)}; \boldsymbol{x}_2^{(0)}]$. Notice that eq. (3.54a) is a ODE since $\boldsymbol{\Lambda}^{(0)}$ is invertible, on the contrary eq. (3.54b) is a pure algebraic equation. Therefore in these coordinates the differential and the algebraic components of the DAE are decoupled. We now differentiate with respect to time eq. (3.54b) and we obtain the following,

$$\boldsymbol{A}_{21}^{(0)}\dot{\boldsymbol{x}}_1^{(0)} + \boldsymbol{A}_{22}^{(0)}\dot{\boldsymbol{x}}_2^{(0)} = \dot{\boldsymbol{f}}_2^{(0)}. \tag{2.55}$$

Writing now eq. (3.55) with eq. (3.54a) in matrix form we obtain the following,

$$\begin{bmatrix} \boldsymbol{\Lambda}^{(0)} & 0 & \boldsymbol{A}_{11}^{(0)} & \boldsymbol{A}_{12}^{(0)} \\ \boldsymbol{A}_{21}^{(0)} & \boldsymbol{A}_{22}^{(0)} & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{\boldsymbol{x}}_1^{(0)} \\ \dot{\boldsymbol{x}}_2^{(0)} \\ \boldsymbol{x}_1^{(0)} \\ \boldsymbol{x}_2^{(0)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1^{(0)} \\ \dot{\boldsymbol{f}}_2^{(0)} \end{bmatrix}. \tag{2.56}$$

As we can observe the differentiation w.r.t time of the algebraic constraints in the matrix representation produces just a shift to the left of the two blocks $\boldsymbol{A}_{21}^{(0)}$, $\boldsymbol{A}_{22}^{(0)}$. Notice also this shift increases the rank of the left block as it is clearly stated below,

$$\operatorname{rank} \begin{bmatrix} \boldsymbol{\Lambda}^{(0)} & 0 \\ 0 & 0 \end{bmatrix} < \operatorname{rank} \begin{bmatrix} \boldsymbol{\Lambda}^{(0)} & 0 \\ \boldsymbol{A}_{\boldsymbol{21}}^{(0)} & \boldsymbol{A}_{\boldsymbol{22}}^{(0)} \end{bmatrix} \tag{2.57}$$

Now we can apply again theorem 9 and find a pair of matrices $\boldsymbol{L}^{(1)} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{U}^{(1)} \in \mathbb{R}^{n\times n}$ such that,

$$\boldsymbol{U}^{(1)} \begin{bmatrix} \boldsymbol{\Lambda}^{(0)} & 0 \\ \boldsymbol{A}_{21}^{(0)} & \boldsymbol{A}_{22}^{(0)} \end{bmatrix} \boldsymbol{L}^{(1)} = \boldsymbol{E}^{(1)} = \begin{bmatrix} \boldsymbol{\Lambda}^{(1)} & 0 \\ 0 & 0 \end{bmatrix},$$

and perform another change of coordinates as follows,

$$\begin{bmatrix} \dot{\boldsymbol{x}}^{(1)} \\ \boldsymbol{x}^{(1)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{U}^{(1)} & 0 \\ 0 & \boldsymbol{I} \end{bmatrix}^{-1} \begin{bmatrix} \dot{\boldsymbol{x}}^{(0)} \\ \boldsymbol{x}^{(0)} \end{bmatrix},$$
$$\boldsymbol{A}^{(1)} = \boldsymbol{L}^{(1)} \boldsymbol{A}^{(0)},$$
$$\boldsymbol{f}^{(1)} = \boldsymbol{L}^{(0)} \begin{bmatrix} \boldsymbol{f}_1^{(0)} \\ \dot{\boldsymbol{f}}_2^{(0)} \end{bmatrix}.$$

Please notice that the definition of $\boldsymbol{f}^{(1)}$ contains part of the field $\boldsymbol{f}^{(0)}$ but also some derivatives of it. Now we can left multiply eq. (3.56) by $\boldsymbol{L}^{(1)}$ and use the change of coordinates defined in eq. (3.58). So finally we reach the following form,

$$\begin{bmatrix} \boldsymbol{E}^{(1)} & \boldsymbol{A}^{(1)} \end{bmatrix} \begin{bmatrix} \dot{\boldsymbol{x}}^{(1)} \\ \boldsymbol{x}^{(1)} \end{bmatrix} = \boldsymbol{f}^{(1)}. \tag{2.59}$$

Here matrix $\begin{bmatrix} \boldsymbol{E}^{(1)} & \boldsymbol{A}^{(1)} \end{bmatrix}$ as seen before has the following structure,

$$\begin{bmatrix} \boldsymbol{E}^{(1)} & \boldsymbol{A}^{(1)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Lambda}^{(1)} & 0 & \boldsymbol{A}_{11}^{(1)} & \boldsymbol{A}_{12}^{(1)} \\ 0 & 0 & \boldsymbol{A}_{21}^{(1)} & \boldsymbol{A}_{22}^{(1)} \end{bmatrix}. \tag{2.60}$$

It is important to point out that the dimensions of matrix $\boldsymbol{\Lambda}^{(1)}$ are greater than the ones of matrix $\boldsymbol{\Lambda}^{(0)}$, formally

$$\dim \boldsymbol{\Lambda}^{(1)} > \dim \boldsymbol{\Lambda}^{(0)}. \tag{2.61}$$

It is therefore clear that at each iteration (derivation) of the constraints the dimensions of $\boldsymbol{\Lambda}^{(k)}$, for $k = 1, \ldots, \nu$ increases until $\dim \boldsymbol{\Lambda}^{(\nu)} = n$ and the matrix $\begin{bmatrix} \boldsymbol{E}^{(\nu)} & \boldsymbol{A}^{(\nu)} \end{bmatrix}$ coincide with $\begin{bmatrix} \boldsymbol{\Lambda}^{(\nu)} & \boldsymbol{A}^{(\nu)} \end{bmatrix}$. At this point the matrix $\boldsymbol{\Lambda}^{(\nu)} \in \mathbb{R}^{n\times n}$ is full rank and then the DAE can be easily transformed into an ODE.

$$\boldsymbol{\Lambda}^{(\nu)} \dot{\boldsymbol{x}}^{(\nu)} + \boldsymbol{A}^{(\nu)} \boldsymbol{x}^{(\nu)} = \boldsymbol{f}^{(\nu)}. \tag{2.62}$$

The minimum integer such that $\dim \boldsymbol{\Lambda}^{(\nu)} = n$ is the Kronecker index (or differential index) of the DAE.

Example 6. Consider the following set of linear differential equation,

$$\dot{x}_1(t) + \dot{x}_2(t) + \dot{x}_3(t) = 2x_1(t) + x_3(t) + t, \tag{2.63a}$$
$$\dot{x}_1(t) + \dot{x}_2(t) + \dot{x}_3(t) = x_1(t) + x_2(t) + t^3, \tag{2.63b}$$
$$0 = x_1(t) + x_2(t) + x_3(t) + \sin t, \tag{2.63c}$$

that we can write in a more compact matrix form as follows,

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \dot{x}_3(t) \end{bmatrix} = \begin{bmatrix} 2 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix} + \begin{bmatrix} t \\ t^3 \\ \sin t \end{bmatrix}. \tag{2.64}$$

Which corresponds to the usual form $E\dot{x}(t) = Ax(t) + f(t)$ obvious meaning for the symbols. Now consider to form the matrix $\begin{bmatrix} E & A & f(t) \end{bmatrix} \in \mathbb{R}^{n \times 2n+1}$ in such a way that,

$$\begin{bmatrix} E & A & f(t) \end{bmatrix} \begin{bmatrix} \dot{x}(t) \\ -x(t) \\ -1 \end{bmatrix} = 0 \tag{2.65}$$

First it is necessary to check if the pair (E, A) is a regular pencil,

$$\det(A - \lambda_0 E) = \det \begin{bmatrix} 2 - \lambda_0 & -\lambda_0 & 1 - \lambda_0 \\ 1 - \lambda_0 & 1 - \lambda_0 & -\lambda_0 \\ 1 & 1 & 1 \end{bmatrix} = 2 \tag{2.66}$$

the determinant is different from zero, so the pair (E, A) is a regular pencil.

Now it is easy to manipulate the following equations

$$\begin{bmatrix} 1 & 1 & 1 & 2 & 0 & 1 & t \\ 1 & 1 & 1 & 1 & 1 & 0 & t^3 \\ 0 & 0 & 0 & 1 & 1 & 1 & \sin t \end{bmatrix} \tag{2.67}$$

∘

Chapter 3

Differential Algebraic Equations for Mechanics

3.1 Overview of Lagrangian Mechanics

Most of the following section is inspired by [1] and [7], which are advanced classical textbooks on classical mechanics. Mechanics deals with the dynamics of particles, rigid bodies, continuous media and field theories such as electromagnetism, gravity, etc. The foundations were laid by Newton in the 17th century and since then many generalizations and extensions have been proposed. To date Mechanics has two main point of view, *Lagrangian mechanics* and *Hamiltonian mechanics*. In these notes we will briefly recall some concepts of Lagrangian Mechanics, for insights see [7]. The Lagrangian formulation of mechanics is based on the observation that there are *variational principles* behind the fundamental laws of force balance, as given by Newton's law $\boldsymbol{F} = m\boldsymbol{a}$. This principle is called *least action*, or more accurately, the principle of *stationary action*. This principle can be successfully applied to every mechanical system to obtain the equations of motion. The idea is fairly general, since it can be used to derive Newtonian, Lagrangian, and Hamiltonian equations of motion. However, here we will focus only on mechanical systems with a finite number of *degrees of freedom*. For those systems the *configuration* can be identified through a vector of *generalized coordinates* $\boldsymbol{q} := (q_1; q_2; \dots; q_n) \in \mathbb{R}^n$, where n is the *minimum* number of *degrees of freedom*. Notice that the choice of vector $\boldsymbol{q}$ is not unique. After choosing a ordered set of coordinate $\boldsymbol{q}$, it is always possible to express *kinetic* and the *potential* energy as a function of it. Thus let us define $T(\boldsymbol{q}, \boldsymbol{q})$ as the *kinetic energy* and $U(\boldsymbol{q}, t)$ as the *potential energy*. Roughly we can think T, U as functions whose domain of definition is simply the direct product of Euclidean spaces, i.e., $T : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ and $U : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$. However this is not completely true, since vector $\boldsymbol{q}$ may represent a coordinatization of more abstract surfaces, namely *manifolds*; however (locally) we can think at Euclidean spaces. Once the *energy* of the system has been defined we can introduce the so called *Lagrangian function* $L(\boldsymbol{q}, \dot{\boldsymbol{q}}, t)$, whose domain is $L : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$. In classical Mechanics the Lagrangian is defined as the difference between *kinetic* and *potential* energy, formally

$$L(\boldsymbol{q}, \dot{\boldsymbol{q}}, t) := T(\boldsymbol{q}, \dot{\boldsymbol{q}}) - U(\boldsymbol{q}, t). \tag{3.1}$$

Example 7. Let us consider a very simple example, a free mass point in the plane subject to gravity in the y downward direction, see example 1. The kinetic and the potential energy for the system are,

$$T = \frac{1}{2} m(\dot{x}^2 + \dot{y}^2), \quad U = mgy,$$

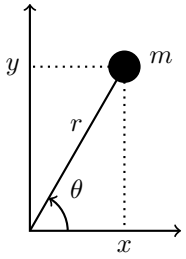

Figure 3.1: A free point of mass m in the plane with gravity.

thus taking $\boldsymbol{q} = (x; y) \in \mathbb{R}^2$ as a vector of generalized coordinates the Lagrangian function is of the form

$$L = \frac{1}{2} m \dot{\boldsymbol{q}}^\intercal \dot{\boldsymbol{q}} - mg\boldsymbol{e}^\intercal \boldsymbol{q},$$

where $e_2 = (0; 1) \in \mathbb{R}^2$. Nevertheless Cartesian coordinates is not the only possible choice, thus considering *polar* coordinates the following relations hold

$$\begin{aligned} x &= r\cos\theta, & y &= r\sin\theta, \\ \dot{x} &= \dot{r}\cos\theta - r\sin\theta\dot{\theta}, & \dot{y} &= \dot{r}\sin\theta + r\cos\theta\dot{\theta}, \end{aligned}$$

where $r \in \mathbb{R}_{\geq 0}$ is the radius and $\theta \in [0, 2\pi)$ the angular coordinate. Then taking some steps the kinetic and potential energies can be expressed as,

$$T = \frac{1}{2} m(\dot{r}^2 + r^2\dot{\theta}^2), \quad U = mgr\sin\theta.$$

Thus taking $\widetilde{\boldsymbol{q}} := (\widetilde{q}_1; \widetilde{q}_2) = (r; \theta) \in \mathbb{R}^2$ as generalized coordinates the Lagrangian is

$$L = \frac{1}{2} m(\dot{\widetilde{q}}_1^2 + \widetilde{q}_1^2\dot{\widetilde{q}}_2^2) - mg\widetilde{q}_1\sin\widetilde{q}_2. \tag{3.3}$$

This simple example shows that the choice of generalized coordinates is not unique but it can be performed according to a criterion of simplicity or utility. As we will see shortly the Lagrangian formulation provides the correct equation of motion despite the chosen reference system. ∘

A very related concept in classical mechanics is the *action* $S \in \mathbb{R}$, this is defined as the integral of the Lagrangian function performed between two instants of time $t_1, t_2 \in \mathbb{R}$, i.e.,

$$S := \int_{t_1}^{t_2} L(\boldsymbol{q}, \dot{\boldsymbol{q}}, t)\,\mathrm{d}t. \tag{3.4}$$

Technically speaking the action is a *functional*, i.e., a map that associate to each element of a vector space (usually a space of curves) a scalar. Thus a functional takes curves as input an returns a scalar.

Remark 5. Notice that $S = S(\boldsymbol{q})$ should be thought as a function of $\boldsymbol{q}(t)$ over the whole time interval $[t_1, t_2]$. From this point onwards we will use the notation $\boldsymbol{q}(\cdot)$ to denote the function $\boldsymbol{q}$, while $\boldsymbol{q}(t)$ will be denote the *value of function* $\boldsymbol{q}$ at the point t. Clarified this it is very important to stress that S does not depends on $q(t)$ but from $q(\cdot)$ and it is just a number. Thus for every choice of the "shape" of curve $\boldsymbol{q}(\cdot)$, the action S assumes a different (scalar) value. ⌟

Example 8. Suppose that $\boldsymbol{q} = (q_1; q_2) \in \mathbb{R}^2$, $t_1 = 0$, $t_2 = 1$ and $L = q_1^2 + q_2^2$, thus the action is the following,

$$S = \int_0^1 q_1^2 + q_2^2\,\mathrm{d}t.$$

Take now $\boldsymbol{q} = (2; t)$, calculating the value for the action results

$$S = \int_0^1 2^2 + t^2 \,\mathrm{d}t = \left[4t + \frac{t^3}{3}\right]_0^1 = \frac{13}{3}.$$

∘

Similarly as seen in analysis courses you may ask yourself if there exist a curve that minimizes or maximizes the action. The answer has created a new branch of mathematics, i.e., the *calculus of variations*. Calculus of variations concerns the *extremal* of functionals whose domain is an infinite-dimensional space: the space of curves. As we will see shortly the *stationary action principle* simply says that every mechanical system evolves along a curve $\boldsymbol{q}(\cdot)$ that is an extremal of the action functional S. In order to state more formally the stationary action principle, let us consider an initial time t_1 and an initial configuration associated to it, $\boldsymbol{q}_1 = \boldsymbol{q}(t_1)$, consider also a final time t_2 and a final configuration $\boldsymbol{q}_2 = \boldsymbol{q}(t_2)$.

Theorem 7 (Stationary action)**.** *Motions of mechanical systems from configuration $\boldsymbol{q}_1$ to configuration $\boldsymbol{q}_2$, with $t \in [t_1, t_2]$, coincide with extremals curves of the functional in eq.* (2.4)*. Or equivalently, for small variations of the curve $\boldsymbol{q}(\cdot)$ the action S is stationary, i.e.,*

$$\delta S = 0. \tag{3.5}$$

Where δ denotes a small perturbation of the functional S and sometimes it is also called *variational operator*.

Example 9. The δ operator can be successfully applied also to ordinary calculus. Let us consider the function $f(x, y) = \cos(x) + y^2$, then a variation of f results

$$\delta f = \frac{\partial f}{\partial x}\delta x + \frac{\partial f}{\partial y}\delta y = \begin{pmatrix}\cos x & y\end{pmatrix}\begin{pmatrix}\delta x \\ \delta y\end{pmatrix}.$$

In this setting stationary points are the ones that satisfies $\delta f = 0$ since for *first order* variation $(\delta x, \delta y)$ the corresponding variation of f is zero. (The surface is locally flat and the variation on f is visible only at the second order). Therefore the stationary points are $(\frac{\pi}{2} + k\pi, 0)$ for $k = 1, 2, \ldots$. ∘

Now let us develop the consequences of this principle. Assuming L regular enough is possible to move the δ operator under the integral,

$$\delta S = \int_{t_1}^{t_2} \delta L(\boldsymbol{q}, \dot{\boldsymbol{q}}, t)\,\mathrm{d}t = \int_{t_1}^{t_2} \frac{\partial L}{\partial \boldsymbol{q}}\delta\boldsymbol{q} + \frac{\partial L}{\partial \dot{\boldsymbol{q}}}\delta\dot{\boldsymbol{q}}\,\mathrm{d}t. \tag{3.6}$$

Remark 6. Notice that variations are performed at *fixed time*, therefore the generic variation $\delta L(\boldsymbol{q}(t), \dot{\boldsymbol{q}}(t), t) = \frac{\partial L}{\partial \boldsymbol{q}}\delta\boldsymbol{q} + \frac{\partial L}{\partial \dot{\boldsymbol{q}}}\delta\dot{\boldsymbol{q}}$ and not $\delta L(\boldsymbol{q}(t), \dot{\boldsymbol{q}}(t), t) \neq \frac{\partial L}{\partial \boldsymbol{q}}\delta\boldsymbol{q} + \frac{\partial L}{\partial \dot{\boldsymbol{q}}}\delta\dot{\boldsymbol{q}} + \frac{\partial L}{\partial t}\delta t$. Moreover we must think at the $\dot{\boldsymbol{q}}$ variable as independent of $\boldsymbol{q}$, since it can be proven that for small variations those quantities are uncorrelated each other. ⌟

Now it is useful to consider the following equality that directly follows from integration by part,

$$\frac{\partial L}{\partial \dot{\boldsymbol{q}}}\delta\dot{\boldsymbol{q}} = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{\boldsymbol{q}}}\delta\boldsymbol{q}\right) - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{\boldsymbol{q}}}\delta\boldsymbol{q}, \tag{3.7}$$

then substituting eq. (2.7) into eq. (2.6) results,

$$\delta S = \int_{t_1}^{t_2} \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{\boldsymbol{q}}}\delta\boldsymbol{q}\right) - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \boldsymbol{q}}\delta\boldsymbol{q} + \frac{\partial L}{\partial \boldsymbol{q}}\delta\boldsymbol{q}\,\mathrm{d}t \tag{3.8a}$$

$$= \left[\frac{\partial L}{\partial \dot{\boldsymbol{q}}}\delta\boldsymbol{q}\right]_{t_1}^{t_2} + \int_{t_1}^{t_2}\left(\frac{\partial L}{\partial \boldsymbol{q}} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{\boldsymbol{q}}}\right)\delta\boldsymbol{q}\,\mathrm{d}t \tag{3.8b}$$

$$= \int_{t_1}^{t_2}\left(\frac{\partial L}{\partial \boldsymbol{q}} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{\boldsymbol{q}}}\right)\delta\boldsymbol{q}\,\mathrm{d}t = 0 \tag{3.8c}$$

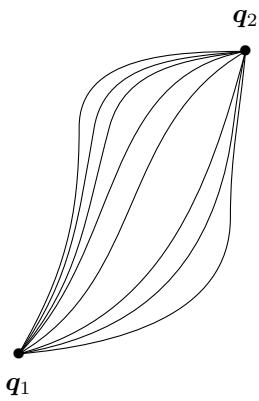

Figure 3.2: A graphical illustration of the possible variations with fixed endpoints $\boldsymbol{q}_1, \boldsymbol{q}_2$.

Remark 7. In developing passage (2.8b) we used $\left[\frac{\partial L}{\partial \dot{\boldsymbol{q}}}\delta\boldsymbol{q}\right]_{t_1}^{t_2} = 0$, since variations $\delta\boldsymbol{q}(t_1) = \delta\boldsymbol{q}(t_2)$ are assumed to be zero because we fixed *a priori* initial and final conditions $\boldsymbol{q}_1$,$\boldsymbol{q}_2$. The concept is graphically illustrated in fig. (2.2), where a set of possible variations is considered but all are vanishing at the endpoints. ⌟

We are now ready to further simplify eq. (2.8b) using the *fundamental lemma of calculus of variations*, which adapted for our purposes states the following:

Lemma 7.3 (Fundamental lemma of calculus of variations)**.** *Let S be a* differentiable *(in variational sense) functional, then if*

$$\int_{t_1}^{t_2} \left(\frac{\partial L}{\partial \boldsymbol{q}} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{\boldsymbol{q}}}\right)\delta\boldsymbol{q}\,\mathrm{d}t = 0, \tag{3.9}$$

for any smooth variation $\delta\boldsymbol{q}$ with $\delta\boldsymbol{q}_1 = \delta\boldsymbol{q}_2 = 0$, then,

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{\boldsymbol{q}}}^{\intercal} - \frac{\partial L}{\partial \boldsymbol{q}}^{\intercal} = 0. \tag{3.10}$$

The eq. (2.10) are called *Euler-Lagrange* equations and they provide the equations of motions for mechanical systems. Roughly speaking lemma 4.3 states that since the variation $\delta\boldsymbol{q}$ is arbitrary, the only way to zeroing the action's variation δS is to satisfy Euler-Lagrange equations. Thus we can collect our results in the following theorem.

Theorem 8. *The curve $q : t \mapsto q(t)$, for $t \in [t_1, t_2]$, is an extremal of the functional S on the space of curves passing through the points $\boldsymbol{q}(t_1) = \boldsymbol{q}_1$ and $\boldsymbol{q}(t_2) = \boldsymbol{q}_2$, if and only if*

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{\boldsymbol{q}}}^{\intercal} - \frac{\partial L}{\partial \boldsymbol{q}}^{\intercal} = 0. \tag{3.11}$$

The key idea of theorem 5 is that a curve $\boldsymbol{q}(\cdot)$ in order to be a stationary point for the action S, must satisfy the Euler-Lagrange equations, and therefore according to theorem 4 Euler-Lagrange equations provide the equations of motion for mechanical systems. Notice also that the variational approach is *coordinate free*, meaning that, Euler-Lagrange equations describe the motion regardless the chosen set of coordinates. Indeed the condition for a curve to be an extremal of a functional does not depend on the choice of coordinate systems.

Example 10. Considering the free mass point in example 1 the equations of motions in the Cartesian coordinates are

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{\boldsymbol{q}}}^{\intercal} - \frac{\partial L}{\partial \boldsymbol{q}}^{\intercal} = m\ddot{\boldsymbol{q}} + mg\boldsymbol{e}_2 = 0, \tag{3.12}$$

while in polar coordinates,

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{\widetilde{q}}_1} - \frac{\partial L}{\partial \widetilde{q}_1} = m\ddot{\widetilde{q}}_1 + mg\sin\widetilde{q}_2, \tag{3.13a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{\widetilde{q}}_2} - \frac{\partial L}{\partial \widetilde{q}_2} = m\widetilde{q}_1^2\ddot{\widetilde{q}}_2 + 2m\widetilde{q}_1\dot{\widetilde{q}}_1\dot{\widetilde{q}}_2 + mg\widetilde{q}_1\cos\widetilde{q}_2 = 0. \tag{3.13b}$$

Solutions $q(\cdot)$ and $\widetilde{q}(\cdot)$ respectively solutions of eq. (2.12) and eq. (2.13) describe the same motion in the plane, just in different coordinate systems. ∘

Nevertheless as presented so far the Euler-Lagrange approach has a big limitation, since L depends only on kinetic and potential energy, we can describe only systems subject to forces that admit a potential function U, i.e., *conservative systems*. Luckily the Euler-Lagrange approach can be extended to *nonconservative forces* using the *D'Alembert's principle*, thus eq. (2.10) is modified as follows

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L^\intercal}{\partial \dot{\boldsymbol{q}}} - \frac{\partial L^\intercal}{\partial \boldsymbol{q}} = \boldsymbol{\tau}, \tag{3.14}$$

where $\boldsymbol{\tau} \in \mathbb{R}^n$ is a vector of *generalized* non conservative forces.

3.2 Mechanical equations of motion

Although Euler-Lagrange equations are very general is often useful to consider an explicit representation for mechanical systems with some special properties. For this purpose we will assume that the kinetic energy $T(\boldsymbol{q}, \dot{\boldsymbol{q}})$ has a very special structure, i.e., is a quadratic form in the generalized velocities $\dot{\boldsymbol{q}}$. This is not a very restrictive assumption since every mechanical system with a finite number of degrees of freedom owns a quadratic kinetic energy. From this simple assumption a lot of interesting properties can be derived; for a good survey see [10]. An explicit expression for kinetic energy is

$$T(\boldsymbol{q}, \dot{\boldsymbol{q}}) = \frac{1}{2}\sum_{j=1}^{n}\sum_{k=1}^{n} m_{jk}(\boldsymbol{q})\dot{q}_j\dot{q}_k = \frac{1}{2}\dot{\boldsymbol{q}}^\intercal \boldsymbol{M}(\boldsymbol{q})\dot{\boldsymbol{q}}, \tag{3.15}$$

where $\boldsymbol{M}(\boldsymbol{q}) \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix, usually called *mass matrix*. The entries $m_{jk}(\boldsymbol{q}) \in \boldsymbol{M}(\boldsymbol{q})$, for $j, k = 1, \ldots, n$ are *generalized masses* and *generalized inertias*.

Remark 8. The matrix $\boldsymbol{M}(\boldsymbol{q})$ is symmetric since the quadratic form associated to a matrix only depends on symmetric part. ⌟

We are now interested in developing an explicit representation for the equations of motion associated to mechanical systems with a *quadratic* kinetic energy. In order to do so we need to perform some tedious calculations. Let us start from Euler-Lagrange equations in eq. (2.10), the term $\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}_i}$ appears as

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}_i} = \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial T}{\partial \dot{q}_i} = \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial}{\partial \dot{q}_i}\left(\frac{1}{2}\sum_{j=1}^{n}\sum_{k=1}^{n} m_{jk}(\boldsymbol{q})\dot{q}_j\dot{q}_k\right).$$

for $i = 1, \ldots, n$. Then exploiting the linearity of summation we can exchange the order between partial derivative and the sum, so that

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}_i} = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}\sum_{j=1}^{n}\sum_{k=1}^{n} m_{jk}(\boldsymbol{q})\frac{\partial \dot{q}_j}{\partial \dot{q}_i}\dot{q}_k + \frac{1}{2}\sum_{j=1}^{n}\sum_{k=1}^{n} m_{jk}(\boldsymbol{q})\dot{q}_j\frac{\partial \dot{q}_k}{\partial \dot{q}_i}\right) \tag{3.16a}$$

$$= \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}\sum_{k=1}^{n} m_{ik}(\boldsymbol{q})\dot{q}_k + \frac{1}{2}\sum_{j=1}^{n} m_{ji}(\boldsymbol{q})\dot{q}_j\right) = \frac{\mathrm{d}}{\mathrm{d}t}\left(\sum_{j=1}^{n} m_{ij}(\boldsymbol{q})\dot{q}_j\right). \tag{3.16b}$$

In last passage we used the fact that $\frac{\partial \dot{q}_j}{\partial \dot{q}_i} = 1$ for $i = j$ and 0 otherwise, and the same holds for $\frac{\partial \dot{q}_k}{\partial \dot{q}_i}$, moreover we also used the symmetry $m_{ij}(\boldsymbol{q}) = m_{ji}(\boldsymbol{q})$ for $i, j = 1, \ldots, n$. Finally deriving with respect to time results,

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}_i} = \sum_{j=1}^{n} m_{ij}(\boldsymbol{q})\ddot{q}_j + \sum_{j=1}^{n}\sum_{k=1}^{n} \frac{\partial m_{ij}(\boldsymbol{q})}{\partial q_k}\dot{q}_k\dot{q}_j. \tag{3.17}$$

The obtained expression in eq. (2.17) is just a part of the Euler-Lagrange equations; considering then the term $\frac{\partial L}{\partial q_i}$, for $i = 1, \ldots, n$, results

$$\frac{\partial L}{\partial q_i} = \frac{1}{2}\sum_{j=1}^{n}\sum_{k=1}^{n} \frac{\partial m_{jk}(\boldsymbol{q})}{\partial q_i}\dot{q}_k\dot{q}_j - \frac{\partial U(\boldsymbol{q}, t)}{\partial q_i}. \tag{3.18}$$

In order to obtain a more compact representation let us define

$$g_i(\boldsymbol{q}) := \frac{\partial U(\boldsymbol{q}, t)}{\partial q_i},$$

for $i = 1, \ldots, n$ and the associate vector $\boldsymbol{g}(\boldsymbol{q}) := (g_1(\boldsymbol{q}); \ldots; g_n(\boldsymbol{q})) \in \mathbb{R}^n$. Combining together eq. (2.17) and eq. (2.18), the overall equations of motion written componentwise result,

$$\sum_{j=1}^{n} m_{ij}(\boldsymbol{q})\ddot{q}_j + \sum_{j=1}^{n}\sum_{k=1}^{n} \frac{\partial m_{ij}(\boldsymbol{q})}{\partial q_k}\dot{q}_j\dot{q}_k - \frac{1}{2}\sum_{j=1}^{n}\sum_{k=1}^{n} \frac{\partial m_{jk}(\boldsymbol{q})}{\partial q_i}\dot{q}_j\dot{q}_k + g_i(\boldsymbol{q}) = \tau_i, \tag{3.19}$$

for $i = 1, \ldots, n$. As it is clear from eq. (2.19) there is a quadratic term in the generalized velocities $\dot{q}_j\dot{q}_k$, therefore it is convenient to introduce the symbol $h_{ijk}(\boldsymbol{q}) \in \mathbb{R}$, with $i, j, k = 1, \ldots, n$, defined as

$$h_{ijk}(\boldsymbol{q}) := \frac{\partial m_{ij}(\boldsymbol{q})}{\partial q_k} - \frac{1}{2}\frac{\partial m_{jk}(\boldsymbol{q})}{\partial q_i}. \tag{3.20}$$

Thus substituting eq. (2.20) into eq. (2.19) the equations of motion result in the more compact representations

$$\sum_{j=1}^{n} m_{ij}(\boldsymbol{q})\ddot{q}_i + \sum_{j=1}^{n}\sum_{k=1}^{n} h_{ijk}(\boldsymbol{q})\dot{q}_j\dot{q}_k + g_i(\boldsymbol{q}) = \tau_i, \tag{3.21}$$

for $i = 1, \ldots, n$. Now let us analyse eq. (2.21) to provide a physical interpretation of the different terms:

- The coefficients $m_{ii}(\boldsymbol{q})$, with $i = 1, \ldots, n$, represent the generalized masses associated to the i-th degree of freedom.
- The coefficients $m_{ij}(\boldsymbol{q})$, with $i, j = 1, \ldots, n$ but $i \neq j$, represent the *inertia forces* induced by the j-th degree of freedom over the i-th.
- The quadratic terms $h_{ijj}\dot{q}_j^2$, with $j = 1, \ldots, n$, represent *centrifugal effect* induced on the i-th degree of freedom by the j-th. Notice that the i-th degree of freedom does not induce any centrifugal effect on itself, i.e., $h_{iii} = 0$.
- The terms $h_{ijj}\dot{q}_j\dot{q}_k$, with $j, k = 1, \ldots, n$ but $i \neq j$ represents the *Coriolis* effects induced on the degree of freedom i-th by velocities of degrees of freedom j-th and k-th.
- The terms $g_i(\boldsymbol{q})$, with $i = 1, \ldots, n$, represent the *conservative generalized forces* acting on the i-th degree of freedom.

At this point it is very useful to rewrite the equations of motion (2.21) in a compact matrix notation as follows:

$$\boldsymbol{M}(\boldsymbol{q})\ddot{\boldsymbol{q}} + \boldsymbol{C}(\boldsymbol{q}, \dot{\boldsymbol{q}})\dot{\boldsymbol{q}} + \boldsymbol{g}(\boldsymbol{q}) = \boldsymbol{\tau}, \tag{3.22}$$

where $\boldsymbol{C}(\boldsymbol{q}, \dot{\boldsymbol{q}})$ is a suitable matrix that takes into account the centrifugal and the Coriolis effects. The choice of this matrix is not unique, since it is sufficient that the elements c_{ij} satisfies the relation

$$\sum_{j=1}^{n} c_{ij}(\boldsymbol{q})\dot{q}_j = \sum_{j=1}^{n}\sum_{k=1}^{n} h_{ijk}(\boldsymbol{q})\dot{q}_j\dot{q}_k. \tag{3.23}$$

Nevertheless a very common choice is to set the elements $c_{ij}(\boldsymbol{q}) \in \boldsymbol{C}(\boldsymbol{q}, \dot{\boldsymbol{q}})$ as

$$c_{ij}(\boldsymbol{q}) = \sum_{k=1}^{n} c_{ijk}(\boldsymbol{q})\dot{q}_k, \tag{3.24}$$

with $i, j = 1, \ldots, n$. Here the symbol $c_{ijk}(\boldsymbol{q}) \in \mathbb{R}$ is called *Christoffel symbols of the first type* and it is defined as

$$c_{ijk}(\boldsymbol{q}) = \frac{1}{2}\left(\frac{\partial m_{ij}(\boldsymbol{q})}{\partial q_k} + \frac{\partial m_{ik}(\boldsymbol{q})}{\partial q_j} - \frac{\partial m_{jk}(\boldsymbol{q})}{\partial q_i}\right). \tag{3.25}$$

Notice that thanks to the symmetry of $\boldsymbol{M}(\boldsymbol{q})$ the following remarkable property of indici exchange holds,

$$c_{ijk}(\boldsymbol{q}) = c_{ikj}(\boldsymbol{q}). \tag{3.26}$$

Moreover, if the Coriolis-centrifugal matrix $C(\boldsymbol{q}, \dot{\boldsymbol{q}})$ is chosen according to eq. (2.24), then the following matrix

$$\dot{\boldsymbol{M}}(\boldsymbol{q}) - 2\boldsymbol{C}(\boldsymbol{q}, \dot{\boldsymbol{q}}), \tag{3.27}$$

is skew symmetric. This property is sometimes useful to synthesize control algorithms for mechanical systems.

3.3 Euler-Lagrange equations with redundant coordinates

What we developed until now holds for systems with a *minimum set of coordinates*. Nevertheless for many real systems is quite hard to write directly kinetic and potential energy using a small set of coordinates. A more common approach is to introduce more generalized coordinates than necessary and then add constraints. This fact provides the motivation to investigate the equations of motions for mechanical systems using a non-minimum set of generalized coordinates. For this purpose let us consider that vector $\boldsymbol{q} \in \mathbb{R}^n$ *does not* represent a minimum set of coordinates. Therefore in order to correctly represent the system we need to impose some constraints. A set of constrain equations can be formally thought as a vectorial function of the form $\boldsymbol{\phi} : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^m$ where m is the number of constrains. The type of constraints equations that we consider are *holomic*, meaning that they depend only on generalized coordinates $\boldsymbol{q}$ and possibly time t. Holomic constrains can be expressed in the following general form,

$$\boldsymbol{\phi}(\boldsymbol{q}, t) = 0. \tag{3.28}$$

Remark 9. The total number of degrees of freedom for the system is not $n-m$, unless some regularity for $\boldsymbol{\phi}(\boldsymbol{q}, t)$ is assumed. Technically if $\forall \boldsymbol{q} \in \mathbb{R}^n$, $\forall t \in \mathbb{R}$ the rank $\frac{\partial \boldsymbol{\phi}(\boldsymbol{q},t)}{\partial \boldsymbol{q}} = m$, then constraints are globally linearly independent each other and the system owns exactly $n - m$ degrees of freedom. Trivially we implicitly assumed that $\boldsymbol{\phi}$ is differentiable. ⌟

Now our aim is to derive the equations of motions for a constrained mechanical systems described by a Lagrangian function $L(\boldsymbol{q}, \dot{\boldsymbol{q}}, t)$ and subject to constrain equation $\boldsymbol{\phi}(\boldsymbol{q}, t) = 0$. In order to do so, we can use the technique of Lagrange multipliers on functional spaces. So let us define the *modified action* $\overline{S} \in \mathbb{R}$ as

$$\overline{S} := \int_{t_1}^{t_2} L(\boldsymbol{q}, \dot{\boldsymbol{q}}, t) + \boldsymbol{\lambda}^\intercal \boldsymbol{\phi}(\boldsymbol{q}, t) \, \mathrm{d}t. \tag{3.29}$$

Here $\lambda \in \mathbb{R}^m$ is a time dependent vector of *Lagrange multipliers*. The following result is fundamental for our purpose.

Theorem 9 (Stationary action for constrained systems)**.** *Motion of mechanical systems from configuration $\boldsymbol{q}_1$ to configuration $\boldsymbol{q}_2$, with $t \in [t_1; t_2]$, coincide with extremals curves of the functional (2.29). Or equivalently, for small variations of the curve $\boldsymbol{q}(\cdot)$ the modified action $\overline{S}$ is stationary, i.e.,*

$$\delta \overline{S} = 0.$$

Remark 10. Notice that we implicitly assumed that initial and final configuration $\boldsymbol{q}_1$, $\boldsymbol{q}_2$ are consistent with constrain equations, i.e., $\boldsymbol{\phi}(\boldsymbol{q}_1, t_1) = \boldsymbol{\phi}(\boldsymbol{q}_2, t_2) = 0$. ⌟

Thus we can think to perform variation as done before for systems with a minimal set of coordinates, but there is a difference, since now $\overline{S}$ also depends on the Lagrange multipliers. Therefore the variational operator δ performs variations on $\boldsymbol{q}$, $\dot{\boldsymbol{q}}$ and $\boldsymbol{\lambda}$. The result is the following,

$$\delta \overline{S} = \int_{t_1}^{t_2} \delta L(\boldsymbol{q}, \dot{\boldsymbol{q}}, t) + \delta \boldsymbol{\lambda}^\intercal \boldsymbol{\phi}(\boldsymbol{q}, t) + \boldsymbol{\lambda}^\intercal \delta \boldsymbol{\phi}(\boldsymbol{q}, t) \, \mathrm{d}t \tag{3.30a}$$

$$= \int_{t_1}^{t_2} \left(\frac{\partial L}{\partial \boldsymbol{q}} - \frac{\mathrm{d}}{\mathrm{d}t} \frac{\partial L}{\partial \dot{\boldsymbol{q}}} + \boldsymbol{\lambda}^\intercal \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}} \right) \delta \boldsymbol{q} + + \delta \boldsymbol{\lambda}^\intercal \boldsymbol{\phi}(\boldsymbol{q}, t) \, \mathrm{d}t + \left[\frac{\partial L}{\partial \dot{\boldsymbol{q}}} \delta \boldsymbol{q} \right]_{t_1}^{t_2} \tag{3.30b}$$

$$= \int_{t_1}^{t_2} \begin{bmatrix} \frac{\partial L}{\partial \boldsymbol{q}} - \frac{\mathrm{d}}{\mathrm{d}t} \frac{\partial L}{\partial \dot{\boldsymbol{q}}} + \boldsymbol{\lambda}^\intercal \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}} & \boldsymbol{\phi}^\intercal \end{bmatrix} \begin{bmatrix} \delta \boldsymbol{q} \\ \delta \boldsymbol{\lambda} \end{bmatrix} \mathrm{d}t = 0, \tag{3.30c}$$

again we used that endpoints variations $\delta \boldsymbol{q}_1 = \delta \boldsymbol{q}_2 = 0$. Finally applying the fundamental lemma of calculus of variations eq. (2.30a), results

$$\frac{\mathrm{d}}{\mathrm{d}t} \frac{\partial L}{\partial \dot{\boldsymbol{q}}}^\intercal - \frac{\partial L}{\partial \boldsymbol{q}}^\intercal = \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}}^\intercal \boldsymbol{\lambda}, \tag{3.31a}$$

$$\boldsymbol{\phi}(\boldsymbol{q}, t) = 0. \tag{3.31b}$$

The equations (2.31) describe the motion of a mechanical system subject to holomic constraints. Note that the term $\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}}^\intercal \boldsymbol{\lambda}$ plays the role of a constrain generalized force which pushes the system to satisfy the constraints $\boldsymbol{\phi}(\boldsymbol{q}, t) = 0$. Unfortunately problem in eq. (2.31) is completely different from the one in eq. (2.10), since it is a mixture of *differential* and *algebraic* equations, formally a *Differential Algebraic Equation* (DAE). Moreover, apparently eq. (2.31), does not provide any way to determine the value of the Lagrange multipliers. Similarly in presence of external non-conservative generalized forces the equation of motions for a constrained mechanical system are modified as follows,

$$\frac{\mathrm{d}}{\mathrm{d}t} \frac{\partial L}{\partial \dot{\boldsymbol{q}}}^\intercal - \frac{\partial L}{\partial \boldsymbol{q}}^\intercal = \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}}^\intercal \boldsymbol{\lambda} + \boldsymbol{\tau}, \tag{3.32a}$$

$$\boldsymbol{\phi}(\boldsymbol{q}, t) = 0, \tag{3.32b}$$

again with $\tau \in \mathbb{R}^n$ a vector of generalized non-conservative forces.

Example 11. Let us consider a simple mechanical example, the pendulum, see example 5. Pendulum own one-degree of freedom and taking the angular coordinate $\theta \in [0, 2\pi)$ as

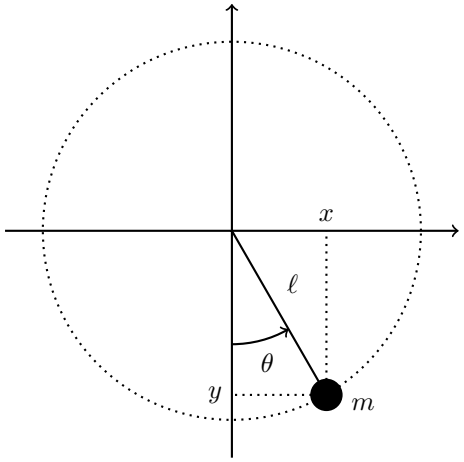

Figure 3.3: A simple pendulum.

generalized coordinate, the following relationships with $(x, y) \in \mathbb{R}^2$ hold,

$$x = \ell \sin\theta, \quad y = -\ell \cos\theta.$$

Thus kinetic and potential energy can be expressed as,

$$T(\theta) = \frac{1}{2}m(\dot{x}^2 + \dot{y}^2) = \frac{1}{2}m\ell^2\dot{\theta}^2, \quad U(\theta) = gmy = -gm\ell\cos\theta,$$

while the Lagrangian for the system results,

$$L = \frac{1}{2}m\ell^2\dot{\theta}^2 + gm\ell\cos\theta.$$

Using the Euler-Lagrange equations in eq. (2.10) yields to

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{\theta}} - \frac{\partial L}{\partial \theta} = m\ell^2\ddot{\theta} + mg\ell\sin\theta = 0,$$

which can be simplified to

$$\ddot{\theta} = -\frac{g}{\ell}\sin\theta.$$

Now suppose we want to describe the pendulum using a set of redundant coordinates, e.g., *natural coordinates*. For this aim let us define the generalized coordinate $\boldsymbol{q} = (x; y) \in \mathbb{R}^2$. The associated Lagrangian function takes the form,

$$L = \frac{1}{2}m\dot{\boldsymbol{q}}^\intercal\dot{\boldsymbol{q}} - mg\boldsymbol{e}_2^\intercal\boldsymbol{q},$$

where $\boldsymbol{e}_2 \in \mathbb{R}^2$ is the second vector of the natural basis, roughly speaking the vector $\boldsymbol{e}_2 = (0; 1) \in \mathbb{R}^2$. Obviously now we need to introduce some constraint equation in order to force mass m to slide over the circle, i.e.,

$$\phi(\boldsymbol{q}) = x^2 + y^2 - \ell^2 = \boldsymbol{q}^\intercal\boldsymbol{q} - \ell^2 = 0.$$

Using then eq. (2.31) the resultant system of equations is the following,

$$\begin{aligned} m\ddot{\boldsymbol{q}} + mg\boldsymbol{e}_2 &= 2\boldsymbol{q}\lambda, \\ \boldsymbol{q}^\intercal\boldsymbol{q} - \ell^2 &= 0. \end{aligned}$$

∘

Chapter 4

Resolution and Numerical Integration of Differential Algebraic Equations

As well pointed out in [8] both from theoretical point of view that the numerical, DAEs are much more challenging than ordinary differential equations (ODEs). Some DAEs can be solved using numerical methods for stiff systems but others cannot and they are very sensitive to the step size and this may cause large errors in the solution and numerical instability. For DAEs coming from constrained mechanical systems a fairly general form is the following,

$$\boldsymbol{M}(\boldsymbol{q})\ddot{\boldsymbol{q}} - \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}}^{\intercal} \lambda = \boldsymbol{n}(\boldsymbol{q}, \dot{\boldsymbol{q}}, t), \tag{4.1a}$$

$$\boldsymbol{\phi}(\boldsymbol{q}, t) = 0 \tag{4.1b}$$

where the term $\boldsymbol{n}(\boldsymbol{q}, \dot{\boldsymbol{q}}, t) \in \mathbb{R}^n$ collects all the contributions coming from Coriolis and centrifugal effects, conservative and non-conservative forces, e.g., it may be definite as follows,

$$\boldsymbol{n}(\boldsymbol{q}, \dot{\boldsymbol{q}}, t) = \boldsymbol{\tau} - \boldsymbol{C}(\boldsymbol{q}, \dot{\boldsymbol{q}})\dot{\boldsymbol{q}} - \boldsymbol{g}(\boldsymbol{q}).$$

It is well known that the equations of motion for a mechanical system are a system of second order (possibly) non-linear differential equations. However for clarity and for sake of generality it is useful to consider a *first order representation*. For this purpose let us define $\boldsymbol{p} \in \mathbb{R}^n$ as *generalized velocities*, i.e.,

$$\boldsymbol{p} := \dot{\boldsymbol{q}}. \tag{4.2}$$

Thus substituting eq. (2.35) into eq. (2.34) results,

$$\dot{\boldsymbol{q}} = \boldsymbol{p}, \tag{4.3a}$$

$$\boldsymbol{M}(\boldsymbol{q})\dot{\boldsymbol{p}} - \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}}^{\intercal} \boldsymbol{\lambda} = \boldsymbol{n}(\boldsymbol{q}, \boldsymbol{p}, t), \tag{4.3b}$$

$$\boldsymbol{\phi}(\boldsymbol{q}, t) = 0. \tag{4.3c}$$

Notice that the vector $(\boldsymbol{q}; \boldsymbol{p}) \in \mathbb{R}^{2n}$ represents the *state* of a mechanical system, namely the pair *generalized position* and *generalized velocity*.

Now a good question is how to solve the problem in eq. (2.36), the issue is certainly how to determine the value for the Lagrange multiplier $\boldsymbol{\lambda}$. One possible approach is to transform the problem into a system of pure differential equations. In order to do so we need to differentiate the constraint $\boldsymbol{\phi}(\boldsymbol{q}, t)$, this yields to

$$\frac{\mathrm{d}\boldsymbol{\phi}}{\mathrm{d}t} = \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}}\boldsymbol{p} + \frac{\partial \boldsymbol{\phi}}{\partial t} = 0. \tag{4.4}$$

Notice that when constraint does not depend on time eq. (2.37) expresses the intuitive idea that generalized velocity $\boldsymbol{p}$ must be orthogonal to constraint's gradient. This provide also the opportunity to notice that constraint $\boldsymbol{\phi}(\boldsymbol{q}, t)$ imposes relationships also at the velocity and acceleration level, e.g., eq. (2.37) restricts the set of feasible velocities. The constraints can we can obtain differentiating the constraint are somehow called *hidden constraints*. Finally notice that eq. (2.37) is still algebraic in the $\boldsymbol{p}$ coordinate, thus in order to obtain a differential relationship let us differentiate again,

$$\frac{\mathrm{d}^2\boldsymbol{\phi}(\boldsymbol{q}, t)}{\mathrm{d}t^2} = \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}}\boldsymbol{p} + \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}}\dot{\boldsymbol{p}} + \frac{\partial^2 \boldsymbol{\phi}}{\partial t^2} = 0. \tag{4.5}$$

Notice that now eq. (2.38) does not impose any algebraic constraint on the state of the mechanical system $(\boldsymbol{q}; \boldsymbol{p})$. Thus substituting the constraint $\boldsymbol{\phi}(\boldsymbol{q}, t) = 0$ with the one in eq. (2.38) results the following modified problem,

$$\dot{\boldsymbol{q}} = \boldsymbol{p}, \tag{4.6a}$$

$$\boldsymbol{M}(\boldsymbol{q})\dot{\boldsymbol{p}} - \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}}^{\mathsf{T}}\boldsymbol{\lambda} = \boldsymbol{n}(\boldsymbol{q}, \boldsymbol{p}, t), \tag{4.6b}$$

$$-\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}}\dot{\boldsymbol{p}} = \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}}\boldsymbol{p} + \frac{\partial^2 \boldsymbol{\phi}}{\partial t^2} := \boldsymbol{m}(\boldsymbol{q}, \boldsymbol{p}, t). \tag{4.6c}$$

Here to simplify the notation we introduced the symbol $\boldsymbol{m}(\boldsymbol{q}, \boldsymbol{p}, t) \in \mathbb{R}^m$. However even if the idea of differentiating the constraint seem good, the problem in eq. (2.31) and the one in eq. (2.39) are *not equivalent*. This difference produces the so called *drift effect* that will be investigated in next section. At the moment let us ignore this complication and continue on the taken path. System of equations in eq. (2.39) can be also written in compact matrix notation as,

$$\begin{bmatrix} \boldsymbol{I} & 0 & 0 \\ 0 & \boldsymbol{M}(\boldsymbol{q}) & -\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}}^{\mathsf{T}} \\ 0 & -\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}} & 0 \end{bmatrix} \begin{bmatrix} \dot{\boldsymbol{q}} \\ \dot{\boldsymbol{p}} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \boldsymbol{p} \\ \boldsymbol{n}(\boldsymbol{q}, \boldsymbol{p}, t) \\ \boldsymbol{m}(\boldsymbol{q}, \boldsymbol{p}, t) \end{bmatrix}. \tag{4.7}$$

Now it's easy to see that if the matrix on the left hand side of eq. (2.40) is invertible, then we can express the system in explicit form and use this representation to perform numerical simulations. It easy to see that the matrix is invertible if and only if the sub-block,

$$\begin{bmatrix} \boldsymbol{M}(\boldsymbol{q}) & -\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}}^{\mathsf{T}} \\ -\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}} & 0 \end{bmatrix} \in \mathbb{R}^{(n+m)\times(n+m)},$$

is non-singular (full rank). In order to prove this we can use the rank additivity formula which states the following,

$$\operatorname{rank} \begin{bmatrix} \boldsymbol{M}(\boldsymbol{q}) & -\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}}^{\mathsf{T}} \\ -\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}} & 0 \end{bmatrix} = \operatorname{rank} \boldsymbol{M}(\boldsymbol{q}) + \operatorname{rank} \left(\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}} \boldsymbol{M}(\boldsymbol{q})^{-1} \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}}^{\mathsf{T}} \right).$$

First recall that matrix $\boldsymbol{M}(\boldsymbol{q})$ is positive definite (so also non-singular), then $\forall \boldsymbol{q} \in \mathbb{R}^n, \operatorname{rank} \boldsymbol{M}(\boldsymbol{q}) = n$. Thus what is missing is the condition

$$\operatorname{rank} \left(\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}} \boldsymbol{M}(\boldsymbol{q})^{-1} \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}}^{\mathsf{T}} \right) = m, \tag{4.8}$$

which is not generally true. Fortunately assuming the constraints (locally) linearly independent each other, i.e., $\forall \boldsymbol{q} \in \mathbb{R}^n, \forall t$

$$\operatorname{rank} \left(\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}} \right) = m, \tag{4.9}$$

eq. (2.41) holds, matrix in eq. (2.40) is invertible and the overall system can be expressed in *explicit first order form* as follows,

$$\begin{bmatrix} \dot{\boldsymbol{q}} \\ \dot{\boldsymbol{p}} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & 0 & 0 \\ 0 & \boldsymbol{M}(\boldsymbol{q}) & -\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}}^{\intercal} \\ 0 & -\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}} & 0 \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{p} \\ \boldsymbol{n}(\boldsymbol{q}, \boldsymbol{p}, t) \\ \boldsymbol{m}(\boldsymbol{q}, \boldsymbol{p}, t) \end{bmatrix}.$$

An explicit expression for the inverse matrix above can be obtained through tedious calculations and it is shown in appendix A.1. However regardless the specific expression of the solution we want stress a very important point; differentiating the constraint we were able to find an explicit expression for the Lagrange multipliers. For this purpose we differentiated twice and we obtained eq. (2.38). Then isolating the acceleration $\dot{\boldsymbol{p}}$ from eq. (2.36) results,

$$\dot{\boldsymbol{p}} = \boldsymbol{M}(\boldsymbol{q}) \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}}^{\intercal} \boldsymbol{\lambda} + \boldsymbol{M}(\boldsymbol{q})^{-1} \boldsymbol{n}(\boldsymbol{q}, \boldsymbol{p}, t) = 0. \tag{4.10}$$

Thus substituting eq. (2.43) into eq. (2.38) results,

$$\frac{\mathrm{d}}{\mathrm{d}t} \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}} \boldsymbol{p} + \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}} \boldsymbol{M}(\boldsymbol{q})^{-1} \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}}^{\intercal} \boldsymbol{\lambda} + \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}} \boldsymbol{M}(\boldsymbol{q})^{-1} \boldsymbol{n}(\boldsymbol{q}, \boldsymbol{p}, t) + \frac{\partial^2 \boldsymbol{\phi}}{\partial t^2}. \tag{4.11}$$

Here we recognize the nonsingular matrix in eq. (2.41), therefore an explicit expression for the Lagrange multipliers is the following,

$$\boldsymbol{\lambda} = -\left(\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}} \boldsymbol{M}(\boldsymbol{q})^{-1} \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}}^{\intercal} \right)^{-1} \left(\frac{\mathrm{d}}{\mathrm{d}t} \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}} \boldsymbol{p} + \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}} \boldsymbol{M}(\boldsymbol{q})^{-1} \boldsymbol{n}(\boldsymbol{q}, \boldsymbol{p}, t) + \frac{\partial^2 \boldsymbol{\phi}}{\partial t^2} \right).$$

Technically Mechanical systems subject to holonomic constraints are systems of differential-algebraic equations (DAEs) of *differential index* 3. The differential index is one plus the number of differentiations of the constraint that are needed in order to be able to eliminate the Lagrange multipliers. Otherwise, equivalently, the number of times that we need to differentiate the constraint in order to obtain a differential equation also for the Lagrange multiplier.

4.1 The drift effect

In this section we want investigate the effects of the differentiating process done to transform problem in eq. (2.36) into problem eq. (2.39). Unfortunately this method is *not* priceless, indeed formally we substitute an algebraic constraint with its second derivative w.r.t time. To better understand the consequences let us consider a fictitious constraint $\widetilde{\boldsymbol{\phi}}(\boldsymbol{q}, t)$ of the form,

$$\widetilde{\boldsymbol{\phi}}(\boldsymbol{q}, t) := \boldsymbol{\phi}(\boldsymbol{q}, t) + \boldsymbol{a}_0 + \boldsymbol{a}_1 t. \tag{4.12}$$

It is easy to see that consider a mechanical system subject to constraint $\phi(\boldsymbol{q}, t)$ or $\widetilde{\boldsymbol{\phi}}(\boldsymbol{q}, t)$ differentiating twice is exactly the same, i.e.,

$$\frac{\mathrm{d}^2 \boldsymbol{\phi}}{\mathrm{d}t^2} = \frac{\mathrm{d}^2 \widetilde{\boldsymbol{\phi}}}{\mathrm{d}t^2}. \tag{4.13}$$

This fact produces the *drift effect*, i.e., the original constraint $\boldsymbol{\phi}(\boldsymbol{q}, t)$ is violated with an error growing with time. This error grows in the best case linearly, but it can be faster due to *numerical* errors. The reason of this behaviour is the term $\boldsymbol{a}_1 t$ which does not have any effect in the formulation (2.39). Note also that this is not an undesired effect coming from not enough precision of the numerical scheme, but it is intrinsic effect due to substitution of the constraint $\boldsymbol{\phi}$ with its second derivative w.r.t time. A simple solution to this undesired effect is the Baumgarte's stabilization that we will discuss in the next section.

Example 12. Continuing example 5 we provide a first order representation for the pendulum; setting $\boldsymbol{p} = \dot{\boldsymbol{q}}$ results,

$$\begin{aligned} \boldsymbol{p} &= \dot{\boldsymbol{q}}, \\ m\dot{\boldsymbol{p}} + mg\boldsymbol{e}_2 &= 2\boldsymbol{q}\lambda, \\ \boldsymbol{q}^\intercal\boldsymbol{q} - \ell^2 &= 0. \end{aligned}$$

Then in order to eliminate the Lagrange multiplier let us differentiate the constraint,

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = 2\boldsymbol{q}^\intercal\boldsymbol{p} = 0.$$

Notice that as expected velocity $\boldsymbol{p}$ and position $\boldsymbol{q}$ are orthogonal each other. Differentiating again we obtain

$$\frac{\mathrm{d}^2\phi}{\mathrm{d}t^2} = 2\dot{\boldsymbol{q}}^\intercal\boldsymbol{p} + 2\boldsymbol{q}^\intercal\dot{\boldsymbol{p}} = 2\boldsymbol{p}^\intercal\boldsymbol{p} + 2\boldsymbol{q}^\intercal\dot{\boldsymbol{p}} = 0.$$

Now we can write the problem in shape of eq. (2.40),

$$\begin{bmatrix} \boldsymbol{I}_2 & 0 & 0 \\ 0 & m\boldsymbol{I}_2 & -2\boldsymbol{q} \\ 0 & -2\boldsymbol{q}^\intercal & 0 \end{bmatrix} \begin{bmatrix} \dot{\boldsymbol{q}} \\ \dot{\boldsymbol{p}} \\ \lambda \end{bmatrix} = \begin{bmatrix} \boldsymbol{p} \\ -mg\boldsymbol{e}_2 \\ 2\boldsymbol{p}^\intercal\boldsymbol{p} \end{bmatrix}, \tag{4.15}$$

∘

where $\boldsymbol{I}_2 \in \mathbb{R}^{2\times 2}$ is the identity matrix of dimension 2. The explicit value for the Lagrange multipliers is the following

$$\lambda = -\frac{m}{2\boldsymbol{q}^\intercal\boldsymbol{q}}\left(\boldsymbol{p}^\intercal\boldsymbol{p} - g\boldsymbol{q}^\intercal\boldsymbol{e}_2\right) = -\frac{m}{2\ell^2}\left(\boldsymbol{p}^\intercal\boldsymbol{p} - gq_2\right). \tag{4.16}$$

To better appreciate the drift effect we can simulate eq. (2.48) using an ODE solver. The results are showed in fig. 2.4. Simulation has been performed for 50 seconds with the following parameters, $g = 9.81$, $m = 2$, $\ell = 2.5$ and initial conditions $\boldsymbol{q}_0 = (0.86; 2.35) \in \mathbb{R}^2$, $\dot{\boldsymbol{q}}_0 = \boldsymbol{p}_0 = (0, 0) \in \mathbb{R}^2$ and $\lambda_0 = 3.69$, which are all consistent with the constraint and with eq. (2.49).

4.2 Baumgarte's stabilization

A very popular way to stabilize the constraint $\boldsymbol{\phi}(\boldsymbol{q}, t) = 0$ and thus avoid the drift effect is using Baumgarte's stabilization, see [3]. This idea is well presented with other techniques in the survey paper [2]. The key idea is to substitute the constraint with a *differential equation* that *asymptotically* satisfies $\boldsymbol{\phi}(\boldsymbol{q}, t) = 0$. Obviously this technique is effective only if you are interested in the *long term* behaviour, since during transients the constraint may be violated. Since constrained mechanical systems require to differentiations to eliminate the Lagrange multipliers, we consider a linear second order stable differential equation of the form,

$$\frac{\mathrm{d}^2 z(t)}{\mathrm{d}t^2} + 2\xi\omega_\mathrm{n}\frac{\mathrm{d}z(t)}{\mathrm{d}t} + \omega_\mathrm{n}^2 z(t) = 0, \tag{4.17}$$

where $z(t) \in \mathbb{R}$. In this form we recognize the well known equation for second order linear mechanical or electrical systems, such as the *RLC circuit* or the *mass spring damper*. In these framework $\xi \in \mathbb{R}_{>0}$ is the *damping ratio* and $\omega_\mathrm{n} \in \mathbb{R}_{>0}$ is the *natural frequency*. Depending on the value of ξ different cases can arise:

- $0 < \xi < 1$: solution z is *underdamped*,
- $\xi = 1$: solution z is *critically damped*,

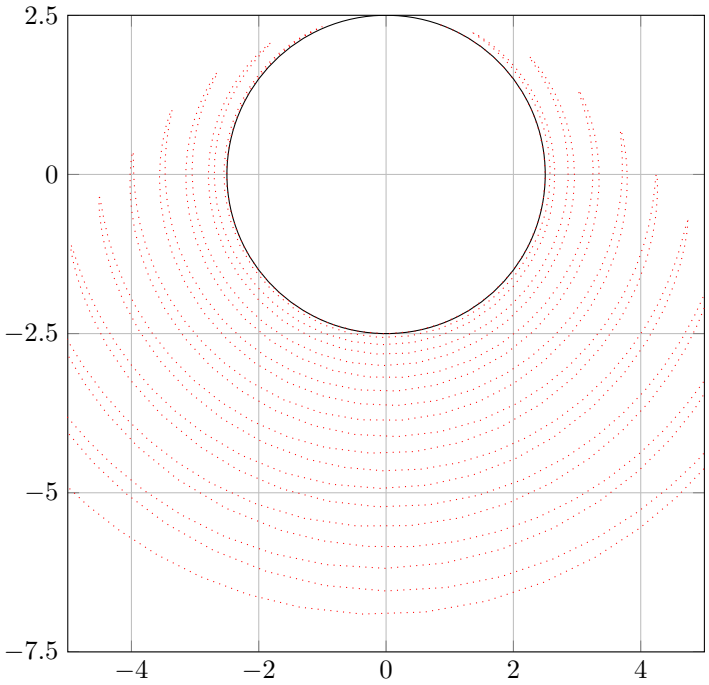

Figure 4.1: Pendulum simulation with drift effect.

- $\xi > 1$: solution z is *overdamped*.

The general form for the solution is the following

$$z(t) = k_1 e^{\left(-\xi+\sqrt{\xi^2-1}\right)\omega_{\mathrm{n}}t} + k_2 e^{\left(-\xi-\sqrt{\xi^2-1}\right)\omega_{\mathrm{n}}t},$$

where $k_1, k_2 \in \mathbb{R}$ are constants determined by the initial conditions $z_0 = z(t=0)$ and $\dot{z}_0 = \dot{z}(t=0)$. Solutions for the different cases can be arranged in different ways, here we show some of them. For a deeper treatment see for example [9]. Let us start from the underdamped case; solution takes the form

$$z(t) = c_1 e^{-\xi\omega_{\mathrm{n}}t} \cos\left(\sqrt{1-\xi^2}\omega_{\mathrm{n}}t - c_2\right), \tag{4.18}$$

where constants $c_1, c_2 \in \mathbb{R}$ are given by,

$$c_1 = \frac{\sqrt{z_0^2\omega_{\mathrm{n}}^2 + \dot{z}_0^2 + 2z_0\dot{z}_0\xi\omega_{\mathrm{n}}}}{\sqrt{1-\xi^2}\omega_{\mathrm{n}}}$$
$$c_2 = \tan^{-1}\left(\frac{\dot{z}_0 + \xi\omega_{\mathrm{n}}z_0}{x_0\omega_{\mathrm{n}}\sqrt{1-\xi^2}}\right).$$

As we can observe in eq. (2.51) solution is a cosine function modulated by a decreasing exponential function. For the critically damped case a linear term appears and the solution takes the form,

$$z(t) = (c_1 + c_2 t)e^{-\omega_{\mathrm{n}}t}, \tag{4.20}$$

where $c_1, c_2 \in \mathbb{R}$ are given by $c_1 = z_0$ and $c_2 = \dot{z}_0 + \omega_{\mathrm{n}}z_0$. Finally for the overdamped case the solution takes the form

$$z(t) = c_1 e^{\left(-\xi+\sqrt{\xi^2-1}\right)\omega_{\mathrm{n}}t} + c_2 e^{\left(-\xi-\sqrt{\xi^2-1}\right)\omega_{\mathrm{n}}t}, \tag{4.21}$$

with constants,

$$c_1 = \frac{x_0\omega_\mathrm{n}\left(\xi + \sqrt{\xi^2-1}\right) + \dot{x}_0}{2\omega_\mathrm{n}\sqrt{\xi^2-1}},$$
$$c_2 = \frac{-x_0\omega_\mathrm{n}\left(\xi - \sqrt{\xi^2-1}\right) - \dot{x}_0}{2\omega_\mathrm{n}\sqrt{\xi^2-1}}.$$

However, regardless of the specific form of the solution, all the three cases satisfy a *global exponential stability* property, i.e., $\forall(z_0, \dot{z}_0) \in \mathbb{R}\times\mathbb{R}$ the solution $z(t, z_0, \dot{z}_0)$ satisfies

$$\lim_{t\to\infty} z(t, z_0, \dot{z}_0) = 0.$$

The notation $z(t, z_0, \dot{z}_0)$ is just to stress that the solution depends on time but also from initial conditions $(z_0, \dot{z}_0)$. This consideration suggests to substitute the constraint $\boldsymbol{\phi}(\boldsymbol{q}, t)$ with an ODE of the form (2.50). In this way, despite the initial conditions, the solution will asymptotically satisfies $\boldsymbol{\phi}(\boldsymbol{q}, t) = 0$. Thus the explicit *Baumgarte's stabilization* for DAEs of index three takes the form,

$$\frac{\mathrm{d}^2\boldsymbol{\phi}}{\mathrm{d}t^2} + 2\xi\omega_\mathrm{n}\frac{\mathrm{d}\boldsymbol{\phi}}{\mathrm{d}t} + \omega_\mathrm{n}^2\boldsymbol{\phi} = 0.$$

Computing explicitly the terms for mechanical systems the stabilization results,

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial\boldsymbol{\phi}}{\partial\boldsymbol{q}}\dot{\boldsymbol{q}} + \frac{\partial\boldsymbol{\phi}}{\partial\boldsymbol{q}}\ddot{\boldsymbol{q}} + \frac{\partial^2\boldsymbol{\phi}}{\partial t^2} + 2\xi\omega_\mathrm{n}\frac{\partial\phi}{\partial\boldsymbol{q}}\dot{\boldsymbol{q}} + 2\xi\omega_\mathrm{n}\frac{\partial\boldsymbol{\phi}}{\partial t} + \omega_\mathrm{n}^2\boldsymbol{\phi} = 0. \tag{4.23}$$

which can be re-arranged in the following form

$$-\frac{\partial\boldsymbol{\phi}}{\partial\boldsymbol{q}}\dot{\boldsymbol{p}} = \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial\boldsymbol{\phi}}{\partial\boldsymbol{q}}\boldsymbol{p} + \frac{\partial^2\boldsymbol{\phi}}{\partial t^2} + 2\xi\omega_\mathrm{n}\frac{\partial\boldsymbol{\phi}}{\partial\boldsymbol{q}}\boldsymbol{p} + 2\xi\omega_\mathrm{n}\frac{\partial\boldsymbol{\phi}}{\partial t} + \omega_\mathrm{n}^2\boldsymbol{\phi} := \boldsymbol{m}(\boldsymbol{q}, \boldsymbol{p}, t). \tag{4.24}$$

Now eq. (2.57) can be easy implemented without any modification in eq. (2.40); indeed it is sufficient to redefine $\boldsymbol{m}(\boldsymbol{q}, \boldsymbol{p}, t)$ as in eq. (2.57). However a question is still open, how can we tune the parameters ω_n and ξ?. Unfortunately there is not general rule, but it is useful to notice that the velocity of convergence to zero of the solution depends on bothξ and ω_n. Therefore it is a good idea to select them in order to obtain an high decay rate compatibly with the *numerical stability*. This means that the dynamics associated to the constraint $\boldsymbol{\phi}(\boldsymbol{q}, t)$ cannot be too fast compared to other ones, otherwise the system could be very hard (stiff) to simulate numerically. Indeed systems where two or more very different time scales are involved are usually called *stiff problems* or *stiff systems*. These may cause some numerical problems, since stiff equations require an integration step extremely small and are numerically unstable. In general it is hard to provide a unique definition of stiffness, but the main idea is that the differential equations includes some terms that can lead to rapid variation in the solution.

Example 13. Continuing example 6 we want use the Baumgarte's technique to stabilize the pendulum. Therefore the idea is to substitute the constraint

$$2\boldsymbol{p}^\mathsf{T}\boldsymbol{p} + 2\boldsymbol{q}\dot{\boldsymbol{p}} = 0$$

with

$$-2\boldsymbol{q}^\mathsf{T}\dot{\boldsymbol{p}} = 2\boldsymbol{p}^\mathsf{T}\boldsymbol{p} + 4\omega_\mathrm{n}\xi\boldsymbol{q}^\mathsf{T}\boldsymbol{p} + \omega_\mathrm{n}^2(\boldsymbol{q}^\mathsf{T}\boldsymbol{q} - \ell^2).$$

Thus final equations appears as,

$$\begin{bmatrix} \boldsymbol{I}_2 & 0 & 0 \\ 0 & m\boldsymbol{I}_2 & -2\boldsymbol{q} \\ 0 & -2\boldsymbol{q}^\mathsf{T} & 0 \end{bmatrix}\begin{bmatrix} \dot{\boldsymbol{q}} \\ \dot{\boldsymbol{p}} \\ \lambda \end{bmatrix} = \begin{bmatrix} \boldsymbol{p} \\ -mg\boldsymbol{e}_2 \\ 2\boldsymbol{p}^\mathsf{T}\boldsymbol{p} + 4\omega_\mathrm{n}\xi\boldsymbol{q}^\mathsf{T}\boldsymbol{p} + \omega_\mathrm{n}^2(\boldsymbol{q}^\mathsf{T}\boldsymbol{q} - \ell^2) \end{bmatrix}. \tag{4.25}$$

Again to appreciate the effects of Baumgarte's stabilization we can simulate eq. (2.58) with same parameters as example 6 and stabilization parameters $\xi = 0.8$, $\omega_\mathrm{n} = 2$. The result is showed in fig. 2.5. ∘

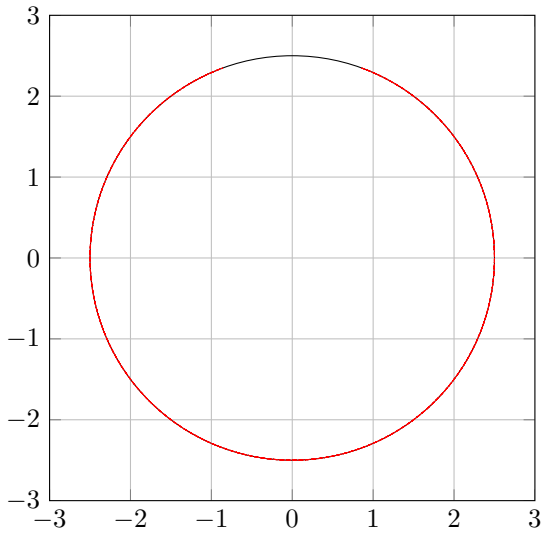

Figure 4.2: Pendulum simulation with Baumgarte's stabilization.

4.3 Projection method

Another way to stabilize the constraint $\boldsymbol{\phi}(\boldsymbol{q}, t) = 0$ and avoid the drift effect is using the so-called Projection method. The idea is to firstly reduce the index of the DAE, thus obtaining an ODE and the aforementioned *hidden constraints*. Then the resulting ODE can be numerically integrated, leading to a generic intermediate solution $\widetilde{\boldsymbol{q}}(t_n)$. This temporary solution is projected to the hidden constraints in order to minimise the error between the numerical integration solution and the actual system constraints, which is a *nonlinear constrained least squares problem* because of the norm chosen. The projection gives the orthogonal projection to the contraints.

It should be noticed that unlike the previously presented Baumgarte's stabilization method this technique is useful and particularly effective if you are interested in the *short term* behaviour, since even during transients the constraint is always satisfied. Different information from the original and reduced system can be used for various projection methods. For example, we can decide to get an advantage by first projecting the ODE solution $\widetilde{\boldsymbol{q}}(t_n)$ to the position constraints and than to the velocity constraints.

Example 14. Continuing example 6 we want use the Projection technique to stabilize the pendulum. The numerical scheme for integrating the reduced DAE is that in eq. (2.48). We want now to project the results of the generic ODE numerical integrator to the following *hidden constraints*,

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = 2\boldsymbol{q}^\intercal \boldsymbol{p} = 0.$$

$$\frac{\mathrm{d}^2\phi}{\mathrm{d}t^2} = 2\dot{\boldsymbol{q}}^\intercal \boldsymbol{p} + 2\boldsymbol{q}^\intercal \dot{\boldsymbol{p}} = 2\boldsymbol{p}^\intercal \boldsymbol{p} + 2\boldsymbol{q}^\intercal \dot{\boldsymbol{p}} = 0.$$

At each time step, the resulting orthogonal projection to the constraints is equal to

$$\begin{aligned} \frac{1}{2}||\widetilde{\boldsymbol{q}} - \boldsymbol{q}||_2 &= \min_{\boldsymbol{q}} \\ \text{subject to:} \quad \boldsymbol{\phi}(\boldsymbol{q}, t) &= \boldsymbol{q}^\intercal \boldsymbol{q} - \ell^2 = 0 \end{aligned} \tag{4.26}$$

for the position coordinates projection, and

$$\begin{aligned} \frac{1}{2}||\widetilde{\boldsymbol{p}} - \boldsymbol{p}||_2 &= \min_{\boldsymbol{p}} \\ \text{subject to:} \quad \frac{\mathrm{d}\boldsymbol{\phi}}{d\boldsymbol{q}}(\boldsymbol{q}, t)\boldsymbol{p} + \frac{\mathrm{d}\boldsymbol{\phi}}{dt}(\boldsymbol{q}, t) &= 0 \end{aligned} \tag{4.27}$$

for the velocity coordinates projection. ◦

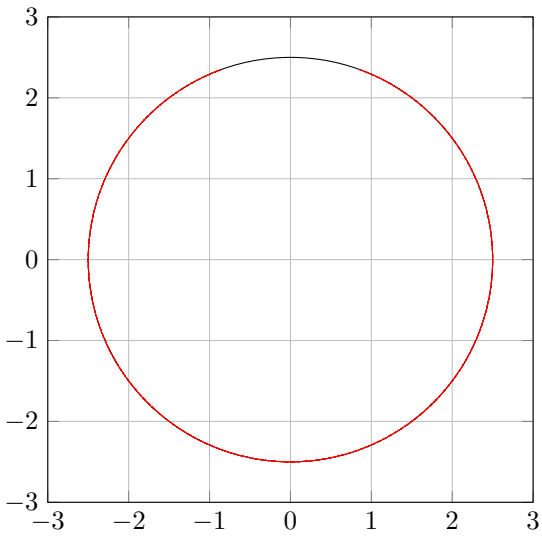

Figure 4.3: Pendulum simulation with Projection method.

Appendix A

A.1 Explicit matrix inverse

Here we provide the explicit expression for the inverse of matrix in eq. (2.40),

$$\begin{bmatrix} \boldsymbol{I} & 0 & 0 \\ 0 & \boldsymbol{M}(\boldsymbol{q}) & -\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}}^{\intercal} \\ 0 & -\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}} & 0 \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{I} & 0 & 0 \\ 0 & \boldsymbol{X}_{11} & \boldsymbol{X}_{12} \\ 0 & \boldsymbol{X}_{21} & \boldsymbol{X}_{22} \end{bmatrix},$$

where the quantities $\boldsymbol{X}_{11} \in \mathbb{R}^{n \times n}$, $\boldsymbol{X}_{12} \in \mathbb{R}^{n \times m}$, $\boldsymbol{X}_{21} \in \mathbb{R}^{m \times n}$, $\boldsymbol{X}_{22} \in \mathbb{R}^{m \times m}$ are defined as follows,

$$\begin{aligned}
\boldsymbol{X}_{11} &= \boldsymbol{M}(\boldsymbol{q})^{-1} - \boldsymbol{M}(\boldsymbol{q})^{-1} \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}}^{\intercal} \left(\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}} \boldsymbol{M}(\boldsymbol{q})^{-1} \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}}^{\intercal} \right)^{-1} \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}} \boldsymbol{M}(\boldsymbol{q})^{-1}, \\
\boldsymbol{X}_{12} &= -\boldsymbol{M}(\boldsymbol{q})^{-1} \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}}^{\intercal} \left(\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}} \boldsymbol{M}(\boldsymbol{q})^{-1} \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}}^{\intercal} \right)^{-1}, \\
\boldsymbol{X}_{21} &= -\left(\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}} \boldsymbol{M}(\boldsymbol{q})^{-1} \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}}^{\intercal} \right)^{-1} \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}} \boldsymbol{M}(\boldsymbol{q})^{-1}, \\
\boldsymbol{X}_{22} &= -\left(\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}} \boldsymbol{M}(\boldsymbol{q})^{-1} \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}}^{\intercal} \right)^{-1}.
\end{aligned}$$

Notice that as expected $\boldsymbol{X}_{12} = \boldsymbol{X}_{21}^{\intercal}$, moreover in the inversion formula appears the expression,

$$\left(\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}} \boldsymbol{M}(\boldsymbol{q})^{-1} \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}}^{\intercal} \right)^{-1}$$

which is exactly the one in eq. (2.41) for which we must guarantee the invertibility. Notice also that in the case of a single constraint, i.e., $\phi(\boldsymbol{q}) \in \mathbb{R}$ the following expression can be simplified as follows,

$$\boldsymbol{M}(\boldsymbol{q})^{-1} \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}}^{\intercal} \left(\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}} \boldsymbol{M}(\boldsymbol{q})^{-1} \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}}^{\intercal} \right)^{-1} \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}} \boldsymbol{M}(\boldsymbol{q})^{-1} = \frac{\boldsymbol{M}(\boldsymbol{q})^{-1} \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}}^{\intercal} \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}} \boldsymbol{M}(\boldsymbol{q})^{-1}}{\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}} \boldsymbol{M}(\boldsymbol{q})^{-1} \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{q}}^{\intercal}}.$$

Bibliography

[1] Vladimir Igorevich Arnold. *Mathematical methods of classical mechanics*, volume 60. Springer Science & Business Media, 1989.

[2] Uri M Ascher, Hongsheng Chin, Linda R Petzold, and Sebastian Reich. Stabilization of constrained mechanical systems with daes and invariant manifolds. *Journal of Structural Mechanics*, 23(2):135–157, 1995.

[3] Joachim Baumgarte. Stabilization of constraints and integrals of motion in dynamical systems. *Computer methods in applied mechanics and engineering*, 1(1):1–16, 1972.

[4] Earl A Coddington and Norman Levinson. *Theory of ordinary differential equations*. Tata McGraw-Hill Education, 1955.

[5] Feliks R Gantmacher. *The theory of matrices*. Taylor & Francis, 1964.

[6] Joao P Hespanha. *Linear systems theory*. Princeton university press, 2009.

[7] Jerrold E Marsden and Tudor Ratiu. *Introduction to mechanics and symmetry: a basic exposition of classical mechanical systems*, volume 17. Springer Science & Business Media, 2013.

[8] Linda Petzold. Differential/algebraic equations are not ode's. *SIAM Journal on Scientific and Statistical Computing*, 3(3):367–384, 1982.

[9] Singiresu S Rao and Fook Fah Yap. *Mechanical vibrations*, volume 4. Addison-Wesley Reading, 1995.

[10] Bruno Siciliano, Lorenzo Sciavicco, Luigi Villani, and Giuseppe Oriolo. *Robotics: modelling, planning and control*. Springer Science & Business Media, 2009.