Commutative algebra in solving differential equations

Debasattam Pal *

1 Introduction

An overwhelming majority of systems dealt with in engineering applications is modelled as differential equations. Irrespective of whether it is a mechanical system, or an electrical system, or a thermal system, or a chemical system, or any hybrid of two or more of such systems, the behavior of such a system is often described using ordinary or partial differential equations (ODEs or PDEs, respectively). This apparent omnipresence of differential equations puts us up against the job of solving such equations, which at times can be quite formidable. One of the standard ways to tackle this problem is to write down the differential equations as first order equations.

The 'order' of an ordinary differential equation is a natural number equal to the highest number of times that the unknown variable is differentiated in the equation. For example, the order of the equation $\frac{\mathrm{d}^3y}{\mathrm{d}t^3} + 6\frac{\mathrm{d}^2y}{\mathrm{d}t^2} + 11\frac{\mathrm{d}y}{\mathrm{d}t} + 6y = 0$ is equal to 3. It had long been realized (more than two hundred years ago) that ODEs that have order equal to 1 are much easier to handle than higher order ones. Let us look at some of the benefits of first order representations.

1. It is easier to solve first order differential equations.
2. Much easier to visualize/pictorially represent the solutions: *velocity vector field*.
3. Physical properties like *stored energy* are easy to represent.
4. Helps in quantifying the *memory* of the system.
5. The well-developed theory of matrices and linear algebra is readily applicable.

Because of the above-mentioned desirable properties, first order representations of systems are always preferred. So much so that often systems are *a priori* assumed to be given in a first order form: called *state-space representation*. However, in many practical scenarios, the mathematical models, obtained by applying various laws of physics, are often higher order differential equations. In this course we shall learn how the mathematical theory of linear and commutative algebra can be applied to carry out the task of obtaining an equivalent first order differential equation from a given higher order differential equation. And then we shall learn how commutative algebraic ideas can be used to solve such equations.

2 The case of a single equation

In this course we shall consider only ordinary linear differential equations with constant coefficients. The simplest case of such differential equations is the one where we have only *one* equation. For example,

$$a_n\frac{\mathrm{d}^n y}{\mathrm{d}t^n} + a_{n-1}\frac{\mathrm{d}^{n-1}y}{\mathrm{d}t^{n-1}} + \cdots + a_1\frac{\mathrm{d}y}{\mathrm{d}t} + a_0 y = 0.$$

*Department of Electrical Engineering, Indian Institute of Technology Bombay. debasattam@ee.iitb.ac.in,

This is the general form of an n^{th} order differential equation. Note that the order n is the smallest positive integer n such that all $a_i = 0$ for $i > n$. Since $a_n \neq 0$, we can divide through all the terms by a_n, and obtain the more standard 'monic' form of the equation as

$$\frac{\mathrm{d}^n y}{\mathrm{d}t^n} + b_{n-1}\frac{\mathrm{d}^{n-1} y}{\mathrm{d}t^{n-1}} + \cdots + b_1\frac{\mathrm{d}y}{\mathrm{d}t} + b_0 y = 0,$$

where $b_i := a_i/a_n$.

We are quite familiar with such type of differential equations from systems and control theory. It turns out to be an easy task to obtain a first order representation for such equations. Indeed, given a monic differential equation

$$\frac{\mathrm{d}^n y}{\mathrm{d}t^n} + a_{n-1}\frac{\mathrm{d}^{n-1} y}{\mathrm{d}t^{n-1}} + \cdots + a_1\frac{\mathrm{d}y}{\mathrm{d}t} + a_0 y = 0,$$

we *create* some new variables like

$$\begin{aligned} x_1 &:= y, \\ x_2 &:= \tfrac{\mathrm{d}y}{\mathrm{d}t}, \\ x_3 &:= \tfrac{\mathrm{d}^2 y}{\mathrm{d}t^2}, \\ &\vdots \\ x_{n-1} &:= \tfrac{\mathrm{d}^{n-2} y}{\mathrm{d}t^{n-2}}, \\ x_n &:= \tfrac{\mathrm{d}^{n-1} y}{\mathrm{d}t^{n-1}}. \end{aligned}$$

It then follows that

$$\begin{aligned} \tfrac{\mathrm{d}x_1}{\mathrm{d}t} &= \tfrac{\mathrm{d}y}{\mathrm{d}t} &&= x_2, \\ \tfrac{\mathrm{d}x_2}{\mathrm{d}t} &= \tfrac{\mathrm{d}^2 y}{\mathrm{d}t^2} &&= x_3, \\ \tfrac{\mathrm{d}x_3}{\mathrm{d}t} &= \tfrac{\mathrm{d}^3 y}{\mathrm{d}t^3} &&= x_4, \\ &\vdots \\ \tfrac{\mathrm{d}x_{n-1}}{\mathrm{d}t} &= \tfrac{\mathrm{d}^{n-1} y}{\mathrm{d}t^{n-1}} &&= x_n, \\ \tfrac{\mathrm{d}x_n}{\mathrm{d}t} &= \tfrac{\mathrm{d}^n y}{\mathrm{d}t^n} &&= -a_{n-1}x_n - \cdots - a_1 x_2 - a_0 x_1. \end{aligned}$$

This is a system of first order equations. Indeed, we can write down the above system of equations in matrix-vector form in the following manner

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-2} & -a_{n-1} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix}.$$

Defining

$$\boldsymbol{x} := \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix}, \text{ and } A := \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-2} & -a_{n-1} \end{bmatrix},$$

we can write down the system of differential equations in a compact form as

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = A\boldsymbol{x}. \tag{1}$$

Equation (1) is the so-called *state-space equation* – well-known in systems and control theory. The vector-valued variable $\boldsymbol{x}$ is called *state variable* and the matrix A is called *system matrix*. The solution to equation (1) is known to be given by the so-called *matrix exponential* in the following manner:

$$\boldsymbol{x}(t) = e^{At}\boldsymbol{x}(0),$$

where $\boldsymbol{x}(0) \in \mathbb{R}^n$ is an n-tuple of real numbers called the *initial condition* vector. The matrix exponential is defined as the following 'convergent' power series of the matrix A:

$$e^{At} := I + At + \frac{A^2t^2}{2!} + \frac{A^3t^3}{3!} + \cdots$$

The solution to the original differential equation is given by the first component of the vector trajectory $\boldsymbol{x}(t)$ because $y(t) = x_1(t)$ follows from definition.

Thus, as mentioned earlier, the case of a single equation is pretty straight-forward. The natural question that arises now is: *what do we do when we have multiple simultaneous equations?* For example, how can we obtain an equivalent first order system of differential equations for a system of differential equations of the form:

$$\begin{array}{rcl} \frac{\mathrm{d}^3y}{\mathrm{d}t^3} + \frac{\mathrm{d}y}{\mathrm{d}t} + y &=& 0, \\ \frac{\mathrm{d}^4y}{\mathrm{d}t^4} + \frac{\mathrm{d}^2y}{\mathrm{d}t^2} + 2y &=& 0. \end{array} \tag{2}$$

Obviously, the method of assigning state-variables do not work out in this case. Then what should be done in this case in order to obtain an equivalent first order system of differential equations? Or, how do we know for sure whether such a representation is possible in the first place?

The answer comes by exploiting the idea of *equivalent system of equations*. We first introduce a notational convention: let ξ be a 'symbol' which is a placeholder for $\frac{\mathrm{d}}{\mathrm{d}t}$ when it is stripped off its operational meaning of differentiation. That is, ξ is a nickname for $\frac{\mathrm{d}}{\mathrm{d}t}$ in its friends' circle outside its work-place. Notice that the two equations in (2) can be written in terms of differential operators as

$$\begin{array}{rcl} \left(\frac{\mathrm{d}^3}{\mathrm{d}t^3} + \frac{\mathrm{d}}{\mathrm{d}t} + 1\right) y &=& 0, \\ \left(\frac{\mathrm{d}^4}{\mathrm{d}t^4} + \frac{\mathrm{d}^2}{\mathrm{d}t^2} + 2\right) y &=& 0. \end{array}$$

Thus, we have two operators $g_1(\xi) = \xi^3 + \xi + 1$ and $g_2(\xi) = \xi^4 + \xi^2 + 2$. Now, we note an interesting fact about the simultaneous equations. Suppose y is a solution to the pair of equations given by (2), and let $f_1(\xi), f_2(\xi)$ be two polynomials with real constant coefficients. Define $g(\xi) := f_1(\xi)g_1(\xi) + f_2(\xi)g_2(\xi)$. Then y also satisfies the differential equation $g\left(\frac{\mathrm{d}}{\mathrm{d}t}\right) y = 0$. This motivates us to do the following. Let us denote by $\mathbb{R}[\xi]$ the set of all polynomials in ξ with constant real coefficients, and let $\mathfrak{C}^\infty(\mathbb{R}, \mathbb{R})$ denote the space of infinitely often differentiable functions from $\mathbb{R}$ to $\mathbb{R}$. Define the following set of polynomials

$$\mathfrak{a} := \{f_1(\xi)g_1(\xi) + f_2(\xi)g_2(\xi) \mid f_1(\xi), f_2(\xi) \in \mathbb{R}[\xi]\}.$$

Then define the following set of trajectories

$$\widetilde{\mathfrak{B}} := \left\{ y \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}) \mid g\left(\frac{\mathrm{d}}{\mathrm{d}t}\right) y = 0 \text{ for all } g(\xi) \in \mathfrak{a} \right\}.$$

Further, let $\mathfrak{B}$ denote the set of smooth solutions to equation (2). From our discussion so far it easily follows that

$$\mathfrak{B} = \widetilde{\mathfrak{B}}.$$

Indeed, $g_1(\xi), g_2(\xi) \in \mathfrak{a}$, and hence any $y \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R})$ that belongs to $\widetilde{\mathfrak{B}}$ also belongs to $\mathfrak{B}$. On the other hand, if $y \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R})$ belongs to $\mathfrak{B}$, as we have seen before, it also belongs to $\widetilde{\mathfrak{B}}$.

Notice that $\widetilde{\mathfrak{B}}$ is described as the solution set of a system of infinitely many differential equations. What purpose does it serve then? With little bit of thinking one would realize that if we are able to find out a polynomial $g(\xi)$ such that $\mathfrak{a} = \{f(\xi)g(\xi) \mid f(\xi) \in \mathbb{R}[\xi]\}$, and denote by $\widetilde{\widetilde{\mathfrak{B}}}$ the solution set of $g\left(\frac{\mathrm{d}}{\mathrm{d}t}\right) y = 0$, then

$$\mathfrak{B} = \widetilde{\mathfrak{B}} = \widetilde{\widetilde{\mathfrak{B}}}.$$

In other words, we would have had a single differential equation that would give us exactly the same set of solutions as that of the given system of simultaneous equations. The question is: when is it possible to obtain such a $g(\xi)$? And if possible then how do we find it? We get answers to these questions from some basic theory in commutative algebra.

3 Commutative algebra: rings and ideals

Commutative algebra is the study of an algebraic object called *commutative ring*.

Definition 1 (Commutative ring)**.** *A* commutative ring with identity *is abstractly defined as a set $\mathcal{A}$ with two* binary operations*, addition $(+)$ and multiplication $(\cdot)$, such that $\mathcal{A}$ satisfies the following properties:*

1. *$\mathcal{A}$ has an additive identity* 0*. That is, $a + 0 = 0 + a = a$ for all $a \in \mathcal{A}$.*
2. *Every element a has an additive inverse $(-a)$. That is, $a + (-a) = (-a) + a = 0$.*
3. *Addition is associative and commutative. That is, for all $a, b, c \in \mathcal{A}$ we have $a + b = b + a$ and $(a + b) + c = a + (b + c)$.*
4. *$\mathcal{A}$ has a multiplicative identity* 1*. That is, $a \cdot 1 = 1 \cdot a = a$ for all $a \in \mathcal{A}$.*
5. *Multiplication is commutative, associative. That is, for all $a, b, c \in \mathcal{A}$ we have $a \cdot b = b \cdot a$ and $(a \cdot b) \cdot c = a \cdot (b \cdot c)$.*
6. *Multiplication distributes over addition. That is, for all $a, b, c \in \mathcal{A}$ we have $a \cdot (b + c) = a \cdot b + a \cdot c$, and $(a + b) \cdot c = a \cdot c + b \cdot c$.*

Example 2. Examples of commutative rings include the set of integers $\mathbb{Z}$, the set of all 1-variable polynomials with constant real (or complex) coefficients $\mathbb{R}[\xi]$ (or $\mathbb{C}[\xi]$). Polynomials in n-variables with real (or complex) coefficients $\mathbb{R}[\xi_1, \ldots, \xi_n]$ (or $\mathbb{C}[\xi_1, \ldots, \xi_n]$).

For notational convenience, we use ab to denote the product $a \cdot b$. An element $a \neq 0$ from a ring $\mathcal{A}$ is said to be a *zero-divisor* if there exists $0 \neq b \in \mathcal{A}$ such that $ab = 0$. A ring that has no zero-divisors in it is called an *integral domain*. All the examples above are examples of integral domains.

A special type of subsets of a ring will be important for us. They are called *ideals*.

Definition 3 (Ideal)**.** *An* ideal *is a subset $\mathfrak{a} \subseteq \mathcal{A}$ such that the following two conditions hold:*

1. *For any two $a, b \in \mathfrak{a}$, $a + b \in \mathfrak{a}$, that is, $\mathfrak{a}$ is* closed under addition*.*
2. *For any $a \in \mathfrak{a}$ and $c \in \mathcal{A}$, $ac \in \mathfrak{a}$, that is, $\mathfrak{a}$ is* closed under multiplication from the ring*.*

Example 4. In the ring of integers $\mathbb{Z}$, consider the set of all even integers $\{0, \pm 2, \pm 4, \pm 6, \ldots\}$. It is easy to check that this set is an ideal. In fact, if $n \in \mathbb{Z}$ then the set of all integers that are multiples of the given integer n, that is, the set $\{0, \pm n, \pm 2n, \pm 3n, \ldots\}$, is an ideal. This ideal is denoted by $n\mathbb{Z}$ or sometimes by $\langle n \rangle$. In this convention, the notation for the ideal of even numbers is $2\mathbb{Z}$ or $\langle 2 \rangle$. It is important to note here that the set of *odd* integers is *not* an ideal; indeed, the set of odd numbers is not closed under addition because the sum of two odd numbers is even.

Example 5. Let $a \in \mathcal{A}$ be given, consider the set $\mathfrak{a} := \{b \in \mathcal{A} \mid$ there exists $q \in \mathcal{A}$ such that $b = qa\}$. Then $\mathfrak{a}$ clearly satisfies the above two conditions; $\mathfrak{a}$ is an ideal. In this case, $\mathfrak{a}$ is said to be *generated by* a, and we write $\mathfrak{a} = \langle a \rangle$. Sometimes, an ideal may be generated by more than one element. In such a case, we write $\mathfrak{a} = \langle a_1, a_2, \ldots, a_r \rangle$, which means

$$\mathfrak{a} := \{q_1 a_1 + q_2 a_2 + \cdots + q_r a_r \mid q_1, q_2, \ldots, q_r \in \mathcal{A}\}.$$

There are rings containing ideals which cannot be generated by finitely many elements. However, because of a result due to D. Hilbert (1862-1943), which is called *Hilbert's Basis Theorem*, it follows that every ideal of the n-variable polynomial ring (for any positive natural number n) is *finitely generated*. An ideal may have multiple sets of generators, which are different from each other. For example, in $\mathbb{Z}$, the ideal $\langle 4, 6 \rangle$ is also generated by 2. Ideals which admit singletons for generating sets are called *principal ideals*. An integral domain where every ideal is principal is called a *principal ideal domain* (PID). For example, $\mathbb{Z}$ is a PID, that is, every ideal in $\mathbb{Z}$ is of the form $\langle n \rangle$ for some $n \in \mathbb{Z}$. Interestingly, and very much crucially for this course, the 1-variable polynomial ring $\mathbb{R}[\xi]$ (also, $\mathbb{C}[\xi]$) is PID, too. This is not so obvious at this point. We shall arrive at this conclusion through the following chain of observations.

We first introduce what is known as the *Euclidean division* process in $\mathbb{R}[\xi]$. Note first that every polynomial in $\mathbb{R}[\xi]$ can be written as

$$f(\xi) = \sum_{i \in \mathbb{N}} a_i \xi^i,$$

where, of course, the sum is finite. Another way of saying it is: only finitely many a_i's are nonzero. Among the non-zero a_i's let n be the highest value of i. This number is of special significance; it is called the *degree* of the polynomial $f(\xi)$, and we denote this number by $\deg(f)$.

Proposition 6 (Euclidean division)**.** *Let* $f(\xi), g(\xi) \in \mathbb{R}[\xi]$. *Suppose* $\deg(g) \leqslant \deg(f)$. *Then there exist* $q(\xi), r(\xi) \in \mathbb{R}[\xi]$ *such that* $f(\xi) = q(\xi)g(\xi) + r(\xi)$ *and* $\deg(r) < \deg(g)$.

Proof : The proof follows from carrying out the long division together with induction on $\deg(f)$. □

Theorem 7. *Let* $\mathfrak{a} \subseteq \mathbb{R}[\xi]$ *be an ideal. Then* $\mathfrak{a}$ *is a principal ideal, that is, there exists* $g(\xi) \in \mathbb{R}[\xi]$ *such that* $\mathfrak{a} = \langle g(\xi) \rangle$. *In other words,* $\mathbb{R}[\xi]$ *is a PID.*

Proof : Let $g(\xi)$ be a non-zero element from the ideal $\mathfrak{a}$ such that $\deg(g) \leqslant \deg(f)$ for all non-zero $f(\xi) \in \mathfrak{a}$. Such a $g(\xi)$ exists because the set of degrees of all non-zero elements in $\mathfrak{a}$ is a subset of $\mathbb{N}$, and $\mathbb{N}$ is bounded from below. We claim that $\langle g(\xi) \rangle = \mathfrak{a}$.
($\langle g(\xi) \rangle \subseteq \mathfrak{a}$) Since $g(\xi) \in \mathfrak{a}$ this inclusion is trivially true.
($\langle g(\xi) \rangle \supseteq \mathfrak{a}$) Let $f(\xi) \in \mathfrak{a}$ be arbitrary. By the choice of $g(\xi)$ we have $\deg(g) \leqslant \deg(f)$. Applying Proposition 6 to this situation we get that there is $q(\xi), r(\xi) \in \mathbb{R}[\xi]$ such that $f(\xi) = q(\xi)g(\xi) + r(\xi)$ and $\deg(r) < \deg(g)$. Now, rearranging the equation $f(\xi) = q(\xi)g(\xi) + r(\xi)$ we get $r(\xi) = f(\xi) - q(\xi)g(\xi)$. Note that $f(\xi) \in \mathfrak{a}$. Further, since $g(\xi) \in \mathfrak{a}$, and $\mathfrak{a}$ is closed under multiplication from $\mathbb{R}[\xi]$, we have $-q(\xi)g(\xi) \in \mathfrak{a}$. Hence $r(\xi) = f(\xi) - q(\xi)g(\xi) \in \mathfrak{a}$ because $\mathfrak{a}$ is closed under addition. Thus, we have a polynomial $r(\xi) \in \mathfrak{a}$ such that $\deg(r) < \deg(g)$. It follows that $r(\xi) = 0$, because if $r(\xi) \neq 0$ than we have a non-zero polynomial $r(\xi) \in \mathfrak{a}$ that has $\deg(r) < \deg(g)$, which is contradictory to the choice of $g(\xi)$ having least possible degree among the elements in $\mathfrak{a}$. However, $r(\xi) = 0$ means $f(\xi) = q(\xi)g(\xi)$, that is, $f(\xi) \in \langle g(\xi) \rangle$. Since $f(\xi) \in \mathfrak{a}$ was chosen arbitrarily it follows that $\langle g(\xi) \rangle \supseteq \mathfrak{a}$. □

This result greatly simplifies the situation. No matter how convoluted the description of an ideal (see Tutorial Question 5 for one such description) might be, it would always simplify as one generated by a single polynomial. However, Theorem 7 does not tell us how to *get*

that generator. Well, this is not exactly true – Theorem 7 does give a method to construct a generator, but, the method requires us to *search* all the polynomials in the ideal to fish out a generator as the one having the least degree. This can be really difficult at times. Fortunately, there is an easier way to construct a generator when we already know a presentation of the given ideal by a finite generating set. That is, suppose we are given with $\{g_1(\xi), g_2(\xi), \ldots, g_m(\xi)\} \subseteq \mathbb{R}[\xi]$, then we can find $g(\xi) \in \mathbb{R}[\xi]$ such that

$$\langle g(\xi)\rangle = \langle g_1(\xi), g_2(\xi), \ldots, g_m(\xi)\rangle.$$

The tool that helps us in achieving this is the idea of a *greatest common divisor (GCD)*. In order to introduce this idea we need the definition of a polynomial being a *divisor* of another polynomial. A polynomial $d(\xi)$ is said to be a *divisor* of a polynomial $f(\xi)$ if there exists a polynomial $q(\xi)$ such that $f(\xi) = q(\xi)d(\xi)$. In this case we write $d(\xi)|f(\xi)$ and say d *divides* f.

Definition 8. *Let $\{g_1(\xi), g_2(\xi), \ldots, g_m(\xi)\} \subseteq \mathbb{R}[\xi]$ be a given set of polynomials. A monic polynomial $g(\xi) \in \mathbb{R}[\xi]$ is said to be a* greatest common divisor (GCD) *of $g_1(\xi), g_2(\xi), \ldots, g_m(\xi)$ if*

1. *$g(\xi)|g_i(\xi)$ for all $i \in \{1, 2, \ldots, m\}$, and*
2. *for every $f(\xi) \in \mathbb{R}[\xi]$ that satisfies $f(\xi)|g_i(\xi)$ for all $i \in \{1, 2, \ldots, m\}$, we must have $f(\xi)|g(\xi)$.*

Proposition 9. *Let $\{g_1(\xi), g_2(\xi), \ldots, g_m(\xi)\} \subseteq \mathbb{R}[\xi]$ be a given set of polynomials. GCD of $g_1(\xi), g_2(\xi), \ldots, g_m(\xi)$ is unique.*

Proof : Let $g(\xi), \widetilde{g}(\xi) \in \mathbb{R}[\xi]$ be two distinct GCDs of $g_1(\xi), g_2(\xi), \ldots, g_m(\xi)$. Then by definition $g(\xi)|\widetilde{g}(\xi)$ and $\widetilde{g}(\xi)|g(\xi)$. Since both $g(\xi)$ and $\widetilde{g}(\xi)$ are monic, this is possible if and only if $g(\xi) = \widetilde{g}(\xi)$. □

This unique GCD is denoted by the notation $\mathrm{GCD}(g_1, g_2, \ldots, g_m)$. The monic GCD can be found out by a neat algorithm known as *Euclidean division algorithm (EDA)*. Before we describe the algorithm, we state and prove the following lemma that will be used crucially in EDA.

Lemma 10. *Let $\{g_1(\xi), g_2(\xi), \ldots, g_m(\xi)\} \subseteq \mathbb{R}[\xi]$ be a given set of polynomials. Suppose $\deg(g_1) \leqslant \deg(g_i)$ for all $i \in \{2, 3, \ldots, m\}$. For $i \in \{2, 3, \ldots, m\}$ let $r_i(\xi)$ be the remainder after division of $g_i(\xi)$ by $g_1(\xi)$. Then*

$$\mathrm{GCD}(g_1, g_2, \ldots, g_m) = \mathrm{GCD}(g_1, r_2, \ldots, r_m).$$

Proof : Define $g(\xi) := \mathrm{GCD}(g_1, g_2, \ldots, g_m)$ and $\widetilde{g}(\xi) := \mathrm{GCD}(g_1, r_2, \ldots, r_m)$. Note that for $i \in \{1, 2, 3, \ldots, m\}$ we have $g(\xi)|g_i(\xi)$. Pick an arbitrary $i \in \{2, 3, \ldots, m\}$. We can write $g_i(\xi) = q_i(\xi)g_1(\xi) + r_i(\xi)$. Since $g(\xi)|g_1(\xi)$ and $g(\xi)|g_i(\xi)$, it follows that $g(\xi)|r_i(\xi)$. Hence, by definition, $g(\xi)|\widetilde{g}(\xi)$ because $\widetilde{g}(\xi) = \mathrm{GCD}(g_1, r_2, \ldots, r_m)$.

On the other hand, for $i \in \{2, 3, \ldots, m\}$, the equation $g_i(\xi) = q_i(\xi)g_1(\xi) + r_i(\xi)$ also implies that $\widetilde{g}(\xi)|g_i(\xi)$ because $\widetilde{g}(\xi)$ divides both $g_1(\xi)$ and $r_i(\xi)$. Once again, by definition of GCD, we must have $\widetilde{g}(\xi)|g(\xi)$ because $g(\xi) = \mathrm{GCD}(g_1, g_2, \ldots, g_m)$.

Thus we have $g(\xi)|\widetilde{g}(\xi)$ and $\widetilde{g}(\xi)|g(\xi)$. Since both $g(\xi)$ and $\widetilde{g}(\xi)$ are monic, this is possible if and only if $g(\xi) = \widetilde{g}(\xi)$. □

Theorem 11 (Euclidean division algorithm (EDA))**.** *Carry out the following algorithm.*

Input: *A finite set of polynomials $\{g_1(\xi), g_2(\xi), \ldots, g_m(\xi)\}$*

Computation:

do

- *Sort the set* $\{g_1(\xi), g_2(\xi), \ldots, g_m(\xi)\}$ *so that* $\deg(g_1) \leqslant \deg(g_i)$ *for all* $i \in \{2, 3, \ldots, m\}$.
- **for** $i \in \{2, 3, \ldots, m\}$
  if $g_i(\xi) \neq 0$
  - *Obtain* $r_i(\xi)$ *by Euclidean division such that* $g_i(\xi) = q_i(\xi)g_1(\xi) + r_i(\xi)$ *with* $\deg(r) < \deg(g_1)$.
  - *Substitute* $g_i(\xi) := r_i(\xi)$.

  end if
  end for

while $g_i(\xi) \neq 0$ *for some* $i \in \{2, 3, \ldots, m\}$

Output *The monic version of* $g_1(\xi)$.

The algorithm stops after finitely many iterations. The output of the algorithm is equal to the $\mathrm{GCD}(g_1, g_2, \ldots, g_m)$

Proof : We first show that the algorithm stops after finitely many passes through the 'do-while' loop. Note that at each pass of the 'do-while' loop, after the sorting step is carried out, the degree of $g_1(\xi)$ strictly decreases. Since $\deg(g_1)$ is lower bounded by 0, the algorithm has to stop after finitely many pass through the 'do-while' loop.

We now show the correctness. When the 'do-while' loop stops, that time only $g_1(\xi) \neq 0$. Now note that, by Lemma 10, it follows that $\mathrm{GCD}(g_1, g_2, \ldots, g_m)$ remains invariant after the reassignment of the polynomials $\{g_1(\xi), g_2(\xi), \ldots, g_m(\xi)\}$ inside the 'do-while' loop. However, when $g_i(\xi) = 0$ for all $i \in \{2, 3, \ldots, m\}$, $\mathrm{GCD}(g_1, g_2, \ldots, g_m)$ is just $g_1(\xi)$. □

The Euclidean division step in the algorithm can be written succinctly using matrix vector notation. Suppose $g_2(\xi)$ is substituted by the remainder $r_2(\xi)$ upon long division by $g_1(\xi)$. We can write this operation as

$$\begin{bmatrix} g_1(\xi) \\ r_2(\xi) \\ g_3(\xi) \\ \vdots \\ g_m(\xi) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -q_2(\xi) & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} g_1(\xi) \\ g_2(\xi) \\ g_3(\xi) \\ \vdots \\ g_m(\xi) \end{bmatrix}.$$

Since, the EDA is nothing but finitely many repeated applications of the Euclidean division step and a permutation step, it is easy to check that the entire process of the algorithm can be encoded as repeated pre-multiplication of the vector $\begin{bmatrix} g_1(\xi) & g_2(\xi) & g_3(\xi) & \cdots & g_m(\xi) \end{bmatrix}^{\mathrm{T}}$ by $m \times m$ matrices having entries from $\mathbb{R}[\xi]$. This repeated pre-multiplication can then be substituted by a single pre-multiplication because the product of these individual matrices is just one single matrix. Thus, we can write

$$\begin{bmatrix} g(\xi) \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = U(\xi) \begin{bmatrix} g_1(\xi) \\ g_2(\xi) \\ g_3(\xi) \\ \vdots \\ g_m(\xi) \end{bmatrix},$$

where $g(\xi) = \mathrm{GCD}(g_1, g_2, \ldots, g_m)$ and $U(\xi) \in \mathbb{R}[\xi]^{m \times m}$. It then clearly follows that, there exists $a_1(\xi), a_2(\xi), \ldots, a_m(\xi)$ such that

$$g(\xi) = a_1(\xi)g_1(\xi) + a_2(\xi)g_2(\xi) + \cdots + a_m(\xi)g_m(\xi).$$

This is known as *Bezout/Aryabhatta identity.*

Theorem 12 (Bezout/Aryabhatta identity)**.** *Let* $\{g_1(\xi), g_2(\xi), \ldots, g_m(\xi)\} \subseteq \mathbb{R}[\xi]$ *be a given set of polynomials. Then there exist* $a_1(\xi), a_2(\xi), \ldots, a_m(\xi) \in \mathbb{R}[\xi]$ *such that*

$$\mathrm{GCD}(g_1, g_2, \ldots, g_m) = a_1(\xi)g_1(\xi) + a_2(\xi)g_2(\xi) + \cdots + a_m(\xi)g_m(\xi). \tag{3}$$

With Bezout/Aryabhatta identity we are now in a position to state and prove the following crucial result of this course.

Theorem 13. *Let* $\{g_1(\xi), g_2(\xi), \ldots, g_m(\xi)\} \subseteq \mathbb{R}[\xi]$ *be a given set of polynomials. Suppose* $g(\xi) = \mathrm{GCD}(g_1, g_2, \ldots, g_m)$*. Then*

$$\langle g_1(\xi), g_2(\xi), \ldots, g_m(\xi)\rangle = \langle g(\xi)\rangle.$$

Proof : $(\langle g_1(\xi), g_2(\xi), \ldots, g_m(\xi)\rangle \subseteq \langle g(\xi)\rangle)$ Let $f(\xi)$ be an arbitrary element in the ideal $\langle g_1(\xi), g_2(\xi), \ldots, g_m(\xi)\rangle$. This means there exist $\alpha_1(\xi), \alpha_2(\xi), \ldots, \alpha_m(\xi) \in \mathbb{R}[\xi]$ such that

$$f(\xi) = \alpha_1(\xi)g_1(\xi) + \alpha_2(\xi)g_2(\xi) + \cdots + \alpha_m(\xi)g_m(\xi). \tag{4}$$

Since $g(\xi)|g_i(\xi)$ for all $i \in \{1, 2, \ldots, m\}$ it follows that every term on the right-hand-side of equation (4) is divisible by $g(\xi)$. Therefore, the entire right-hand-side, and hence $f(\xi)$, is divisible by $g(\xi)$. In other words, $f(\xi) \in \langle g(\xi)\rangle$. Since $f(\xi)$ was chosen arbitrarily, we have $\langle g_1(\xi), g_2(\xi), \ldots, g_m(\xi)\rangle \subseteq \langle g(\xi)\rangle$.
$(\langle g_1(\xi), g_2(\xi), \ldots, g_m(\xi)\rangle \supseteq \langle g(\xi)\rangle)$ It is enough to show that $g(\xi) \in \langle g_1(\xi), g_2(\xi), \ldots, g_m(\xi)\rangle$. This follows directly from Bezout/Aryabhatta identity. Indeed, by Theorem 12 there exist $a_1(\xi), a_2(\xi), \ldots, a_m(\xi) \in \mathbb{R}[\xi]$ such that

$$g(\xi) = a_1(\xi)g_1(\xi) + a_2(\xi)g_2(\xi) + \cdots + a_m(\xi)g_m(\xi) \in \langle g_1(\xi), g_2(\xi), \ldots, g_m(\xi)\rangle.$$

□

4 First order representation of a system of simultaneous scalar ODEs

Suppose we have a system of simultaneous ODEs of the form

$$\begin{array}{rcl} g_1(\frac{\mathrm{d}}{\mathrm{d}t})y &=& 0 \\ g_2(\frac{\mathrm{d}}{\mathrm{d}t})y &=& 0 \\ &\vdots& \\ g_m(\frac{\mathrm{d}}{\mathrm{d}t})y &=& 0. \end{array} \tag{5}$$

Let $\mathfrak{B}$ denote the set of all smooth (that is, $\mathfrak{C}^\infty(\mathbb{R}, \mathbb{R})$) solutions to this given set of equations. Now consider the polynomials $g_1(\xi), g_2(\xi), \ldots, g_m(\xi) \in \mathbb{R}[\xi]$, and define by $\mathfrak{a}$ the ideal generated by these polynomials. That is $\mathfrak{a} := \langle g_1(\xi), g_2(\xi), \ldots, g_m(\xi)\rangle$. Define the following set

$$\mathfrak{B}(\mathfrak{a}) := \left\{ y \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}) \mid f(\frac{\mathrm{d}}{\mathrm{d}t})y = 0 \text{ for all } f(\xi) \in \mathfrak{a} \right\}. \tag{6}$$

It is easy to check the following.

Lemma 14. *Let* $\mathfrak{B}$ *denote the set of all smooth solutions of a given system of simultaneous ODEs as in equation* (5)*. Let* $\mathfrak{a}$ *and* $\mathfrak{B}(\mathfrak{a})$ *be as defined above. Then we have*

$$\mathfrak{B} = \mathfrak{B}(\mathfrak{a}).$$

Proof : Straightforward verification. □

The following main result now is a direct consequence of Lemma 14 and Theorems 7 and 13.

Theorem 15. *Let $\mathfrak{B}$ be the set of all smooth solutions of a given system of simultaneous ODEs*

$$\begin{aligned} g_1(\tfrac{\mathrm{d}}{\mathrm{d}t})y &= 0 \\ g_2(\tfrac{\mathrm{d}}{\mathrm{d}t})y &= 0 \\ &\vdots \\ g_m(\tfrac{\mathrm{d}}{\mathrm{d}t})y &= 0. \end{aligned}$$

Consider the polynomials $g_1(\xi), g_2(\xi), \ldots, g_m(\xi)$ and suppose $g(\xi)$ is the $\mathrm{GCD}(g_1, g_2, \ldots, g_m)$. *Let $\widetilde{\mathfrak{B}}$ be the set of all smooth solutions of*

$$g(\frac{\mathrm{d}}{\mathrm{d}t})y = 0.$$

Then we have

$$\mathfrak{B} = \widetilde{\mathfrak{B}}.$$

Proof : Define the ideal $\mathfrak{a} := \langle g_1(\xi), g_2(\xi), \ldots, g_m(\xi)\rangle$. By Theorem 13 $\mathfrak{a} = \langle g(\xi)\rangle$. Recall the definition of $\mathfrak{B}(\mathfrak{a})$ given by equation (6). By Lemma 14 we first get $\mathfrak{B} = \mathfrak{B}(\mathfrak{a})$. Further, since $\mathfrak{a} = \langle g(\xi)\rangle$, applying Lemma 14 again to this situation we get $\widetilde{\mathfrak{B}} = \mathfrak{B}(\mathfrak{a})$. Thus

$$\mathfrak{B} = \mathfrak{B}(\mathfrak{a}) = \widetilde{\mathfrak{B}}.$$

□

In the remaining part of this course we shall see how remainder finding, Euclidean division etc. can be substituted completely by matrix manipulations and linear equations, and finally, how a linear ODE can be solved using these ideas. This will require us to make use of an altogether different type of algebra: the algebra of remainders.

5 The algebra of remainders AKA the quotient

Suppose we are given a monic polynomial $g(\xi) \in \mathbb{R}[\xi]$. What we now do is we pick *all* possible polynomials and perform Euclidean division on them by $g(\xi)$ and collect the remainders. This collection of *all* remainders has nice algebraic properties. We first notice that, for a given $f(\xi) \in \mathbb{R}[\xi]$, suppose we write $f(\xi) = q(\xi)g(\xi) + r(\xi)$, where $\deg(r) < \deg(g)$, then $r(\xi)$ must be unique. In order to see this, suppose we have two different representations for the same $f(\xi)$ as $f(\xi) = q_1(\xi)g(\xi) + r_1(\xi) = q_2(\xi)g(\xi) + r_2(\xi)$. Rearranging we get that $r_1(\xi) - r_2(\xi) = (q_2(\xi) - q_1(\xi))g(\xi)$. This means $g(\xi)|(r_1(\xi) - r_2(\xi))$, but this is possible only if $(r_1(\xi) - r_2(\xi)) = 0$ because $\deg(r_1 - r_2) \leqslant \max\{\deg(r_1), \deg(r_2)\} < \deg(g)$. Thus, $r_1(\xi) = r_2(\xi)$. Because of this uniqueness, given $f(\xi) \in \mathbb{R}[\xi]$, we can write $[f]$ to denote this unique remainder without any ambiguity. We now define the following set

$$\mathcal{M} := \{[f] \mid f(\xi) \in \mathbb{R}[\xi]\}.$$

Note that several different polynomials may have the same remainder. That is, there could be $f_1(\xi), f_2(\xi) \in \mathbb{R}[\xi]$, $f_1(\xi) \neq f_2(\xi)$ such that $[f_1] = [f_2]$. It is not difficult to check that $[f_1] = [f_2]$ if and only if $g(\xi)|(f_1(\xi) - f_2(\xi))$. Thus, one can define a *relation* on the set $\mathbb{R}[\xi]$ as: $f_1(\xi)$ is related with $f_2(\xi)$ if and only if $[f_1] = [f_2]$ (equivalently, if and only if $g(\xi)|(f(\xi) - f_2(\xi))$). That way, every element in the set $\mathcal{M}$, as defined above, stands for a set of polynomials that are related with each other. Indeed, $[f]$ can be identified with the set $f(\xi) + \langle g(\xi)\rangle$; this set is called the *coset* of $f(\xi)$. This is a standard construction in mathematics known as *factoring* or

quotienting under an *equivalence relation*. The gruesome details about this construction will not be required for this course and is also beyond the limited scope of this course; much much more details of this can be found in textbooks like [AM69, CLO07].

The set $\mathcal{M}$ has the structure of a commutative ring. The addition and multiplication of the elements of $\mathcal{M}$ are not obvious here, they need to be defined. This is done as follows.

Definition 16. *For $f_1(\xi), f_2(\xi) \in \mathbb{R}[\xi]$ we define*

$$[f_1] + [f_2] := [f_1 + f_2],$$

and

$$[f_1][f_2] := [f_1 f_2].$$

Proposition 17. *The addition and multiplication are well-defined.*

Proof : We verify this here only for multiplication and leave the same for addition as an exercise. Suppose $\widetilde{f}_1(\xi), \widetilde{f}_2(\xi) \in \mathbb{R}[\xi]$ are such that $[f_1] = [\widetilde{f}_1]$ and $[f_2] = [\widetilde{f}_2]$. We need to verify that $[f_1 f_2] = [\widetilde{f}_1 \widetilde{f}_2]$. Since $[f_1] = [\widetilde{f}_1]$ we have $f_1(\xi) - \widetilde{f}_1(\xi) = q_1(\xi)g(\xi)$. Hence, $f_1(\xi) = \widetilde{f}_1(\xi) + q_1(\xi)g(\xi)$. Similarly, since $[f_2] = [\widetilde{f}_2]$ we have $f_2(\xi) = \widetilde{f}_2(\xi) + q_2(\xi)g(\xi)$. It then follows that

$$\begin{aligned} f_1(\xi)f_2(\xi) &= \Big(\widetilde{f}_1(\xi) + q_1(\xi)g(\xi)\Big)\Big(\widetilde{f}_2(\xi) + q_2(\xi)g(\xi)\Big) \\ &= \widetilde{f}_1(\xi)\widetilde{f}_2(\xi) + \widetilde{f}_1(\xi)q_2(\xi)g(\xi) + \widetilde{f}_2(\xi)q_1(\xi)g(\xi) + q_1(\xi)q_2(\xi)g(\xi)^2 \\ \Rightarrow f_1(\xi)f_2(\xi) - \widetilde{f}_1(\xi)\widetilde{f}_2(\xi) &= \widetilde{f}_1(\xi)q_2(\xi)g(\xi) + \widetilde{f}_2(\xi)q_1(\xi)g(\xi) + q_1(\xi)q_2(\xi)g(\xi)^2 \\ \Rightarrow [f_1 f_2] &= [\widetilde{f}_1\widetilde{f}_2]. \end{aligned}$$

□

With this notion of addition and multiplication, the set $\mathcal{M}$ becomes a commutative ring (see Tutorial Question 9). In other words, the remainder can be added, subtracted and multiplied; there are additive and multiplicative identities; and there is additive inverse. Note that this also makes $\mathcal{M}$ a finite dimensional vector space. Indeed, if $\deg(g) = n$, then every element in $\mathcal{M}$ would be a polynomial of degree at most $n-1$ because degree of a remainder upon division by $g(\xi)$ is strictly less than $\deg(g)$. Thus, $\{1, \xi, \xi^2, \ldots, \xi^{n-1}\}$ can be taken as a basis for $\mathcal{M}$ as a vector space over $\mathbb{R}$. Note that under this basis, every remainder is identified with a row-vector (of size n) consisting of the coefficients of the remainder in ascending order of their corresponding degrees. For example, the remainder $a_0 + a_1\xi + a_2\xi^2 + \ldots + a_{n-1}\xi^{n-1}$ is identified with $[a_0 \quad a_1 \quad a_2 \quad \cdots \quad a_{n-1}]$. It follows that $\mathcal{M}$ is isomorphic as a vector space with $\mathbb{R}^n$, where $n = \deg(g)$.

Vector spaces are arguably much easier to handle than commutative rings. Therefore, it is tempting to work with $\mathcal{M}$ viewing it as a vector space. However, there is one issue: although, addition and scalar multiplication are exactly the same for $\mathcal{M}$ in both points of view, the vector space structure does not automatically provide with a means to multiply two vectors. Multiplication of vectors can be brought into the picture of $\mathcal{M}$ as a vector space by invoking the notion of linear maps from a vector space to itself. A map $\varphi : \mathcal{M} \to \mathcal{M}$ is said to be *linear* if it satisfies the following two properties:

1. $\varphi([f_1] + [f_2]) = \varphi([f_1]) + \varphi([f_2])$, for all $[f_1], [f_2] \in \mathcal{M}$,
2. $\varphi(\alpha[f]) = \alpha\varphi([f])$, for all $[f] \in \mathcal{M}$ and $\alpha \in \mathbb{R}$.

In what comes next, we shall see that multiplication by a fixed $[f]$ can be viewed as a linear map from $\mathcal{M}$ to itself. Suppose $f(\xi) \in \mathbb{R}[\xi]$ is chosen arbitrarily and fixed. Define the following map $\mu_f : \mathcal{M} \to \mathcal{M}$ in the following manner: for $[h] \in \mathcal{M}$

$$\mu_f([h]) := [fh].$$

It is easy to check that μ_f is linear. Now, once we have a linear map from a finite dimensional vector space to itself, that linear map can be expressed as a square matrix. Indeed, suppose we have

$$\mu_f\left([\xi^i]\right) = a_{i,0} + a_{i,1}\xi + a_{i,2}\xi^2 + \cdots + a_{i,(n-1)}\xi^{n-1} \text{ for } 0 \leqslant i \leqslant n-1. \tag{7}$$

We can store this data in the form of a matrix as shown below

$$A_f := \begin{bmatrix} a_{0,0} & a_{0,1} & a_{0,2} & \cdots & a_{0,(n-1)} \\ a_{1,0} & a_{1,1} & a_{1,2} & \cdots & a_{1,(n-1)} \\ a_{2,0} & a_{2,1} & a_{2,2} & \cdots & a_{2,(n-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{(n-1),0} & a_{(n-1),1} & a_{(n-1),2} & \cdots & a_{(n-1),(n-1)} \end{bmatrix}. \tag{8}$$

An interesting fact emerges from this matrix: the linear map μ_f is completely described by the matrix A_f. In order to see this, take a typical element from $\mathcal{M}$, say $[h]$, and write it in the form a remainder as $[h] = h_0 + h_1\xi + h_2\xi^2 + \cdots + h_{n-1}\xi^{n-1}$, $h_i \in \mathbb{R}$ for all $i \in \{0, 1, 2, \ldots, n-1\}$. Note that, due to linearity of μ_f it follows that

$$\begin{aligned} \mu_f([h]) &= h_0\mu_f([1]) + h_1\mu_f([\xi]) + h_2\mu_f\left(\left[\xi^2\right]\right) + \cdots + h_{n-1}\mu_f\left(\left[\xi^{n-1}\right]\right) \\ &= h_0\left(\sum_{j=0}^{n-1} a_{0,j}\xi^j\right) + h_1\left(\sum_{j=0}^{n-1} a_{1,j}\xi^j\right) + h_2\left(\sum_{j=0}^{n-1} a_{2,j}\xi^j\right) + \cdots + \\ &\qquad\qquad h_{n-1}\left(\sum_{j=0}^{n-1} a_{(n-1),j}\xi^j\right) \\ &= \left(\sum_{i=0}^{n-1} h_i a_{i,0}\right) + \left(\sum_{i=0}^{n-1} h_i a_{i,1}\right)\xi + \left(\sum_{i=0}^{n-1} h_i a_{i,2}\right)\xi^2 + \cdots + \left(\sum_{i=0}^{n-1} h_i a_{i,(n-1)}\right)\xi^{n-1}. \end{aligned}$$

This expression becomes immediately more sensible by noting that if $\mu_f([h])$ is written in the form of a remainder as $b_0 + b_1\xi + b_2\xi^2 + \cdots + b_{n-1}\xi^{n-1}$ then the row-vector of coefficients $[b_0 \quad b_1 \quad b_2 \quad \cdots \quad b_{n-1}]$ satisfies

$$[b_0 \quad b_1 \quad b_2 \quad \cdots \quad b_{n-1}] = [h_0 \quad h_1 \quad h_2 \quad \cdots \quad h_{n-1}]A_f,$$

where $[h_0 \quad h_1 \quad h_2 \quad \cdots \quad h_{n-1}]$ is the row-vector of coefficients corresponding to $[h]$ represented as a remainder. In other words, if we decide to represent the elements of $\mathcal{M}$ by their corresponding row-vectors, say

$$v_h = [h_0 \quad h_1 \quad h_2 \quad \cdots \quad h_{n-1}] \text{ for } [h] = h_0 + h_1\xi + h_2\xi^2 + \cdots + h_{n-1}\xi^{n-1}, \tag{9}$$

then we have

$$v_{\mu_f([h])} = v_{[fh]} = v_h A_f. \tag{10}$$

To express this in one phrase: multiplication by $[f]$ is represented by post-multiplication by A_f.

This observation has far reaching consequences. For example the following proposition.

Proposition 18. *Given a monic polynomial $g(\xi) \in \mathbb{R}[\xi]$ with $\deg(g) = n$, let $\mathcal{M}$ be the corresponding commutative ring formed by the remainders upon division by $g(\xi)$ (the* quotient ring *in short). Further, let $f(\xi), h(\xi) \in \mathbb{R}[\xi]$ be arbitrary, and let $A_f, A_h \in \mathbb{R}^{n\times n}$ be the corresponding matrices as defined by equations* (7) *and* (8)*. Moreover, let $v_f, v_h \in \mathbb{R}^{1\times n}$ be the row-vectors corresponding to representations of $[f], [h] \in \mathcal{M}$ as remainders, respectively. Then the following hold:*

1. *$\mu_{f+h} = \mu_f + \mu_h$ and $A_{f+h} = A_f + A_h$.*

2. $\mu_{fh} = \mu_f \mu_h$ *and* $A_{fh} = A_f A_h$.

3. $v_{fh} = v_h A_f = v_f A_h = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \end{bmatrix} A_{fh}$.

Proof : Straight-forward verification. □

For the special case of $f(\xi) = \xi$, let us use the symbol A to denote the corresponding $A_f \in \mathbb{R}^{n\times n}$. We can actually write down this matrix A explicitly. Suppose $g(\xi) = a_0 + a_1\xi + a_2\xi^2 + \cdots + a_{n-1}\xi^{n-1} + \xi^n$. It is easy to verify that $A \in \mathbb{R}^{n\times n}$ has the following form:

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1} \end{bmatrix}. \tag{11}$$

This matrix is called the *companion matrix*.

Lemma 19. *Let* $g(\xi) \in \mathbb{R}[\xi]$ *be given by* $g(\xi) = a_0 + a_1\xi + a_2\xi^2 + \cdots + a_{n-1}\xi^{n-1} + \xi^n$, *and* $A \in \mathbb{R}^{n\times n}$ *be as defined by equation* (11). *Suppose* $f(\xi) \in \mathbb{R}[\xi]$ *is given by* $f(\xi) = b_0 + b_1\xi + b_2\xi^2 + \cdots + b_m\xi^m$. *Suppose* $v_f \in \mathbb{R}^{1\times n}$ *is the row-vector of the coefficients of* $[f] \in \mathcal{M}$ *represented as a remainder (see equation* (9)*). Then we must have*

$$v_f = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \end{bmatrix} f(A),$$

where

$$f(A) = b_0 I + b_1 A + b_2 A^2 + \cdots + b_m A^m.$$

Further, the following are equivalent:

1. $g(\xi) | f(\xi)$.

2. $f(A) = 0 \in \mathbb{R}^{n\times n}$.

Proof : It follows from Parts 1 and 2 of Proposition 18 that

$$A_f = b_0 I + b_1 A + b_2 A^2 + \cdots + b_m A^m = f(A).$$

Since $v_f \in \mathbb{R}^{1\times n}$ is the row-vector of the coefficients of $[f] \in \mathcal{M}$ represented as a remainder, it follows from Statement 3 of Proposition 18 that

$$v_f = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \end{bmatrix} A_f$$

because $[f] = [f][1] = \mu_f([1])$. This fact together with the fact that $A_f = f(A)$ we get that

$$v_f = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \end{bmatrix} f(A). \tag{12}$$

We now prove the equivalence claimed in the lemma.

(1⇒2) We assume that $g(\xi)|f(\xi)$, and we want to prove that $f(A) = 0$. First note that $g(\xi)|f(\xi)$ implies that $[f] = 0$, that is, $v_f = 0 \in \mathbb{R}^{1\times n}$. It then follows from equation (12) that

$$v_f = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \end{bmatrix} f(A) = 0.$$

Also note that $g(\xi)|f(\xi)$ means that $g(\xi)|\xi^i f(\xi)$ for all $i \in \{1, 2, \ldots, n-1\}$. It then follows that

$$\begin{aligned} v_{\xi^i f} &= \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \end{bmatrix} A^i f(A) = 0 \text{ for all } i \in \{1, 2, \ldots, n-1\} \\ &= \begin{bmatrix} 0 & 0 & \cdots & 1 & \cdots & 0 \end{bmatrix} f(A) = 0, \end{aligned}$$

where the 1 in the row-vector is at the i^{th} position. Taking $i = 1, 2, \ldots, n-1$ one by one it follows that every row of $f(A)$ is the zero row. Thus, $f(A)$ is the zero matrix.

(1⇐2) We assume that $f(A) = 0$. It then follows once again from equation (12) that

$$v_f = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \end{bmatrix} f(A) = 0.$$

Now $v_f = 0$ means the remainder $[f]$ of $f(\xi)$ upon division by $g(\xi)$ is zero, which in turn means that $g(\xi)|f(\xi)$. □

6 Solving linear constant coefficient ODEs using the companion matrix

Once again we come back to linear ODEs with constant real coefficients. As we have seen earlier, every system of equations can be brought down to an equivalent single equation. So, we consider, without loss of generality, that we are given with the following differential equation:

$$g(\frac{\mathrm{d}}{\mathrm{d}t})y = 0$$

where $g(\xi) \in \mathbb{R}[\xi]$ is a monic polynomial of degree n. We notice the following curious fact. Let $y \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R})$ be a solution to the equation $g(\frac{\mathrm{d}}{\mathrm{d}t})y = 0$. Now, suppose we are given $f(\xi) \in \mathbb{R}[\xi]$, and we are asked to find out $f(\frac{\mathrm{d}}{\mathrm{d}t})y$. It turns out that y being a solution of the given differential equation enables us to reduce this question to the following situation. Let $r(\xi)$ be the remainder of $f(\xi)$ upon division by $g(\xi)$. Then we must have

$$f(\frac{\mathrm{d}}{\mathrm{d}t})y = r(\frac{\mathrm{d}}{\mathrm{d}t})y.$$

In order to see why this is true, recall the identity $f(\xi) = q(\xi)g(\xi) + r(\xi)$. It then follows that

$$f(\frac{\mathrm{d}}{\mathrm{d}t})y = q(\frac{\mathrm{d}}{\mathrm{d}t})g(\frac{\mathrm{d}}{\mathrm{d}t})y + r(\frac{\mathrm{d}}{\mathrm{d}t})y.$$

But, $g(\frac{\mathrm{d}}{\mathrm{d}t})y = 0$ because y is a solution. Hence $f(\frac{\mathrm{d}}{\mathrm{d}t})y = r(\frac{\mathrm{d}}{\mathrm{d}t})y$.

This apparently innocuous reduction actually provides us with huge computational saving. Indeed, notice that a general polynomial $f(\xi)$ does not have any restriction on its degree, while the remainder $r(\xi)$ can have degree at most $n-1$. Therefore, if one wants to keep record of the actions of *all* the polynomial differential operators $f(\frac{\mathrm{d}}{\mathrm{d}t})$, it is not required that she takes all $f(\xi) \in \mathbb{R}[\xi]$, it is sufficient that she does so *only* for the remainders $r(\xi)$. Quite interestingly, not only does this reduce the effort of evaluating $f(\frac{\mathrm{d}}{\mathrm{d}t})y$ for solutions y, this also provides a way to *solve* a given differential equation.

Suppose, like before, we have a differential equation $g(\frac{\mathrm{d}}{\mathrm{d}t})y = 0$, where $g(\xi) \in \mathbb{R}[\xi]$ is a monic polynomial of degree n. Let us assume that $y \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R})$ is an exponential solution[1] of the equation $g(\frac{\mathrm{d}}{\mathrm{d}t})y = 0$, that is, y is of the form

$$y(t) := y_0 + y_1 t + y_2 \frac{t^2}{2!} + \cdots + y_k \frac{t^k}{k!} + \cdots,$$

where $y_k \in \mathbb{R}$ for all $k = 1, 2, 3, \ldots$ are such that $y(t) \in \mathbb{R}$ for all $t \in \mathbb{R}$. It then follows that $\left(\frac{\mathrm{d}^k y}{\mathrm{d}t^k}\right)(0) = y_k$. Now note that it follows from our discussion above that

$$\frac{\mathrm{d}^k y}{\mathrm{d}t^k} = \left[\frac{\mathrm{d}^k}{\mathrm{d}t^k}\right] y,$$

where $\left[\xi^k\right]$ is the remainder of ξ^k upon division by $g(\xi)$. This remainder is a polynomial of degree at most $n-1$. That is, $\left[\xi^k\right]$ can be written as a linear combination of $1, \xi, \xi^2, \ldots, \xi^{n-1}$ with real coefficients. Suppose

$$\left[\xi^k\right] = \alpha_0 + \alpha_1 \xi + \cdots + \alpha_{n-1}\xi^{n-1},$$

then it follows that

$$\frac{\mathrm{d}^k y}{\mathrm{d}t^k} = \left[\frac{\mathrm{d}^k}{\mathrm{d}t^k}\right] y = \alpha_0 y + \alpha_1 \frac{\mathrm{d}y}{\mathrm{d}t} + \cdots + \alpha_{n-1}\frac{\mathrm{d}^{n-1}y}{\mathrm{d}t^{n-1}}.$$

[1] This is *not* a restrictive assumption at all. In fact, it can be shown that *every* solution of the given type of differential equations is exponential. However, showing this is beyond the scope of this course.

However, we have already noticed that

$$\left(\frac{\mathrm{d}^k y}{\mathrm{d}t^k}\right)(0) = y_k.$$

Combining these two observations we get that

$$\begin{aligned} y_k = \left(\frac{\mathrm{d}^k y}{\mathrm{d}t^k}\right)(0) &= \left(\left[\frac{\mathrm{d}^k}{\mathrm{d}t^k}\right] y\right)(0) \\ &= \alpha_0 y(0) + \alpha_1 \left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)(0) + \cdots + \alpha_{n-1}\left(\frac{\mathrm{d}^{n-1}y}{\mathrm{d}t^{n-1}}\right)(0) \\ &= \alpha_0 y_0 + \alpha_1 y_1 + \cdots + \alpha_{n-1} y_{n-1}. \end{aligned}$$

Thus, if we know $y_0, y_1, \ldots, y_{n-1}$, we can derive y_k for all $k \geqslant n$. Note that knowing this, we also know the entire solution $y(t)$.

So, finding out the remainders for ξ^k is required in order to know the entire signal. This job of remainder finding can be carried out by making use of the companion matrix. Recall equation (11), where the companion matrix has been derived. From Lemma 19 it follows that the coefficient vector of the remainder representation of $\left[\xi^k\right]$ can be found in the following manner. Suppose $\left[\xi^k\right] = \alpha_0 + \alpha_1\xi + \cdots + \alpha_{n-1}\xi^{n-1}$, then

$$\begin{bmatrix}\alpha_0 & \alpha_1 & \cdots & \alpha_{n-1}\end{bmatrix} = \begin{bmatrix}1 & 0 & \cdots & 0\end{bmatrix} A^k.$$

It then follows that

$$y_k = \begin{bmatrix}\alpha_0 & \alpha_1 & \cdots & \alpha_{n-1}\end{bmatrix} \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{n-1}\end{bmatrix} = \begin{bmatrix}1 & 0 & \cdots & 0\end{bmatrix} A^k \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{n-1}\end{bmatrix}. \tag{13}$$

Denoting the column-vector $\begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{n-1}\end{bmatrix}$ by $\boldsymbol{y}$ we get

$$y_k = \begin{bmatrix}1 & 0 & \cdots & 0\end{bmatrix} A^k \boldsymbol{y}, \text{ for all } k \geqslant n. \tag{14}$$

Using equation (14) we get that the solution can be written as

$$y(t) = \begin{bmatrix}1 & 0 & \cdots & 0\end{bmatrix} \left(\sum_{k=0}^{\infty} \frac{A^k t^k}{k!}\right) \boldsymbol{y}. \tag{15}$$

So far we have assumed that we know that $y(t)$ is an exponential solution, and then we derived that $y(t)$ must satisfy equation (15) in that case. Now suppose that we are given an arbitrary vector $\boldsymbol{x} = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{n-1}\end{bmatrix}$, and we define

$$x(t) := \begin{bmatrix}1 & 0 & \cdots & 0\end{bmatrix} \left(\sum_{k=0}^{\infty} \frac{A^k t^k}{k!}\right) \boldsymbol{x}. \tag{16}$$

Then it is easy to verify that: for $i \in \{1, 2, 3, \ldots\}$

$$\frac{\mathrm{d}^i x}{\mathrm{d}t^i} = \begin{bmatrix}1 & 0 & \cdots & 0\end{bmatrix} \left(\sum_{k=0}^{\infty} \frac{A^k t^{k-i}}{(k-i)!}\right) \boldsymbol{x} = \begin{bmatrix}1 & 0 & \cdots & 0\end{bmatrix} A^i \left(\sum_{k=0}^{\infty} \frac{A^k t^k}{k!}\right) \boldsymbol{x}.$$

It follows that for a polynomial $f(\xi) \in \mathbb{R}[\xi]$ we must have

$$f(\frac{\mathrm{d}}{\mathrm{d}t})x = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} f(A) \left(\sum_{k=0}^{\infty} \frac{A^k t^k}{k!} \right) \boldsymbol{x}.$$

Recall that for any $f(\xi)$ such that $g(\xi)|f(\xi)$ we must have $f(A) = 0$. Therefore, $x(t)$ as defined by equation (16) must satisfy

$$g(\frac{\mathrm{d}}{\mathrm{d}t})x = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} g(A) \left(\sum_{k=0}^{\infty} \frac{A^k t^k}{k!} \right) \boldsymbol{x} = 0$$

because $g(A) = 0$. In other words, $x(t)$ is a solution of the differential equation $g(\frac{\mathrm{d}}{\mathrm{d}t})y = 0$.

7 Tutorial problems

Question 1. Show that the examples given in Example 2 are indeed commutative rings.

Question 2. Show that the set of all smooth functions $\mathfrak{C}^\infty(\mathbb{R},\mathbb{R})$ is a commutative ring. What is the multiplicative identity here? Consider the subset $\mathcal{S}$ of $\mathfrak{C}^\infty(\mathbb{R},\mathbb{R})$ comprising of all smooth functions that are zero over the open interval $(-1,1)$. Show that $\mathcal{S}$ is an ideal. Is this ideal finitely generated?

Question 3. Let $a,b,c \in \mathbb{Z}$. Show that

$$\mathrm{GCD}(a,b,c) = \mathrm{GCD}\left(\mathrm{GCD}(a,b),c\right) = \mathrm{GCD}\left(\mathrm{GCD}(a,c),b\right) = \mathrm{GCD}\left(\mathrm{GCD}(b,c),a\right).$$

Question 4. Prove that $\mathbb{Z}$ is a PID.

Question 5. Consider the following set of polynomials $\{f(\xi) \mid f(-1) = f(1) = f(2) = 0\}$. Show that the set is an ideal. Find out a generator for this ideal.

Question 6. Compute the GCD of the polynomials $\xi^3+6\xi^2+11\xi+6$ and $\xi^4+8\xi^3+21\xi^2+22\xi+8$ by EDA.

Question 7. Show that the only smooth function $y \in \mathfrak{C}^\infty(\mathbb{R},\mathbb{R})$ that satisfies both the equations

$$\begin{aligned} \frac{\mathrm{d}^2y}{\mathrm{d}t^2} + \frac{\mathrm{d}y}{\mathrm{d}t} + 2y &= 0 \\ \frac{\mathrm{d}^3y}{\mathrm{d}t^3} + 7y &= 0 \end{aligned}$$

is the zero function.

Question 8. For the two polynomials in Question 6 find out the two polynomials that will appear in the corresponding Bezout/Aryabhatta identity.

Question 9. Verify that the addition and multiplication defined in $\mathcal{M}$ satisfy all the axioms of commutative ring.

Question 10. Take $g(\xi) = \xi^2+\xi+1$ and obtain remainders of $\xi^3+6\xi^2+11\xi+6$ and ξ^4+1 upon division by $g(\xi)$. Do this once using the companion matrix and then verify it using Euclidean division.

Question 11. Let $g(\xi) \in \mathbb{R}[\xi]$ be monic of degree equal to n, and let $A \in \mathbb{R}^{n\times n}$ be the corresponding companion matrix. Suppose $f(\xi) \in \mathbb{R}[\xi]$. Show that $f(\xi)$ is coprime with $g(\xi)$ (that is, $\mathrm{GCD}(f(\xi),g(\xi)) = 1$) if and only if $f(A) \in \mathbb{R}^{n\times n}$ is invertible.

References

[AM69] M.F. Atiyah and I.G. Macdonald. *Introduction to Commutative Algebra.* Addison-Wesley publishing Co., 1969.

[CLO07] D. Cox, J. Little, and D. O'Shea. *Ideals, Varieties, and Algorithms.* Springer: Undergraduate Texts in Mathematics, 3 edition, 2007.

[Eis95] D. Eisenbud. *Commutative Algebra with a View Toward Algebraic Geometry.* Springer-Verlag, 1995.

[KFA69] R.E. Kalman, P.L. Falb, and M.A. Arbib. *Topics in Mathematical Systems Theory.* McGraw-Hill: International Series in Pure and Applied Mathematics, 1969.

[Pom94] J-F. Pommaret. *Partial Differential Equations and Group Theory:New Perspectives for Applications.* Kluwer Academic Publishers, Dordrecht, 1994.

[PW98] J. W. Polderman and J. C. Willems. *Introduction to Mathematical Systems Theory: A Behavioral Approach*. Springer-Verlag, 1998.

[QP98] A. Quadrat and J-F. Pommaret. Generalized bezout idenity. *Applicable Algebra in Engineering, Communication and Computing*, 9:91–116, 1998.

[Wil91] J.C. Willems. Paradigms and puzzles in the theory of dynamical systems. *IEEE Transactions on Automatic Control*, 36(3):259–294, March 1991.