SJÄLVSTÄNDIGA ARBETEN I MATEMATIK

MATEMATISKA INSTITUTIONEN, STOCKHOLMS UNIVERSITET

Analyzing polynomial dynamical systems using algebraic methods

av

Dennis Öberg

2018 - No M9

MATEMATISKA INSTITUTIONEN, STOCKHOLMS UNIVERSITET, 106 91 STOCKHOLM

Analyzing polynomial dynamical systems using algebraic methods

Dennis Öberg

Självständigt arbete i matematik 30 högskolepoäng, avancerad nivå

Handledare: Yishao Zhou

2018

Abstract

The main purpose of the work presented in this master thesis has been to study how algebraic methods can be used for studying polynomial dynamical systems. Algebraic methods for determining the number of steady states, finding said states and determining the stability properties of them are presented, as well as methods for reducing the dimension of and for reducing the number of parameters in such systems. It is also illustrated how these methods can be used to study some classes of systems which appear in applications.

Sammanfattning

Det huvudsakliga syftet med det arbete som presenteras i denna masteruppsats har varit att studera hur algebraiska metoder kan användas för att studera polynomiella dynamiska system. Algebraiska metoder för att bestämma antalet ekvilibriumpunkter, för att hitta dessa punkter samt avgöra deras stabilitetsegenskaper presenteras, liksom metoder för att reducera dimensionen av och reducera antalet parametrar hos sådana system. Det illustreras också hur dessa metoder kan användas för att studera några klasser av system som förekommer i tillämpningar.

Contents

1 Introduction

When one thinks of methods for analyzing dynamical systems given by systems of ordinary differential equations, what springs to mind is perhaps tools from mathematical analysis. In contrast, the main purpose of the work being presented in this thesis has been to study how algebraic methods can be used for this purpose. More precisely, we investigate the subclass of these systems for which the defining equations are of the form

$$\dot{x}_i = p_i(x_1, x_2, \ldots, x_n),\ \ i = 1, 2, 3, \ldots, n$$

where $p_i \in \mathbb{R}[x_1, x_2, \ldots, x_n]$. We call these polynomial dynamical systems.

As is well-known, we usually can not solve systems of ordinary differential equations explicitly. Rather, we make a qualitative study of such systems. There are several properties which are of interest. Are there any points $x \in \mathbb{R}^n$ such that $p_i(x) = 0$ for all $i = 1, 2, \ldots, n$, i.e. are there any steady states? How can we find the steady states? How does the system behave close to the steady states; in other words, what are the stability properties of the steady states? Also, before starting to analyze a system, it is worthwhile to investigate whether it can be expressed in a simpler form; can the mathematical relations described by the system be expressed using fewer variables and/or parameters? In other words, we want to know if the system can be reduced. We will present a framework for studying these and other properties in the case of polynomial dynamical systems.

It has been a goal of the author to make the presentation accessible to those which encounter polynomial dynamical systems in applications; therefore, the rule of the thumb has been to define explicitly as many of the concepts used as possible. Still, some concepts are assumed to be known, e.g. the concept of a ring.

2 Reduction of the dimension of a polynomial dynamical system

2.1 Introduction

Convention *$\dot{x}$ and $\frac{dx}{dt}$ will be used interchangeably to denote the derivative of $x(t)$.*

Convention *If x is a vector, then $\dot{x}$ (and $\frac{dx}{dt}$) will denote the component-wise derivative.*

Let us start with an example. This system is a common example in the literature; see e.g. [5, chapter 7.1-7.2], [16, chapter 3.2.2], [17, chapter 2.8].

Example 2.1.1 (based on [17, chapter 2.8]). Let E be an enzyme which reacts with a substrate S to form a complex C; temporarily, call this reaction 1. From the complex, a product P is formed and the enzyme E is released — call this reaction 2 — but C also deteriorates back into E and S; call this reaction 3. When biochemists study these systems, they often assume that the law of mass action holds, which means that the rate at which a reaction takes place is proportional to the product of the concentrations of the molecules taking part in the reaction. Let λ be the proportionality constant of reaction 1, κ the proportionality constant of reaction 2 and μ the proportionality constant of reaction 3. Schematically, this can be written

$$\begin{array}{rcl} S+E & \overset{\lambda}{\to} & C \\ C & \overset{\kappa}{\to} & P+E \\ C & \overset{\mu}{\to} & S+E \end{array}.$$

Let $E(t)$ be the concentration of the enzyme, $S(t)$ the concentration of the substrate, $C(t)$ the concentration of the complex and $P(t)$ the concentration of the product at time t. Assume that the law of mass action holds. The dynamics of the concentrations of the different substances are then described by

$$\left\{\begin{array}{rclcr} \dot{S} & = & -\lambda SE & + & \mu C \\ \dot{E} & = & -\lambda SE & + & (\mu+\kappa)C \\ \dot{C} & = & \lambda SE & - & (\mu+\kappa)C \\ \dot{P} & = & & & \kappa C \end{array}\right. . \tag{2.1}$$

The analysis of this system then proceeds by observing that we can add, for example, the second and the third equation to each other, to get $\dot{E}+\dot{C}=0$, which we integrate to get $E(t)+C(t)\equiv a$ (the symbol "$\equiv$" denotes identity), for some $a\in\mathbb{R}$. This is an algebraic relation among the variables of the system, which then is used to express one of the variables in terms of the other. This gives

$$\left\{\begin{array}{rclclcr} \dot{S} & = & -\lambda aS & + & \lambda SC & + & \mu C \\ \dot{C} & = & \lambda aS & - & \lambda SC & - & (\mu+\kappa)C \\ \dot{P} & = & & & & & \kappa C \end{array}\right. .$$

Moreover, $\dot{S}+\dot{E}+\dot{P}=0$, so we also have $S(t)+E(t)+P(t)\equiv b$ for some $b\in R$. This gives

$$\left\{\begin{array}{rclclcr} \dot{S} & = & -\lambda aS & + & \lambda SC & + & \mu C \\ \dot{C} & = & \lambda aS & - & \lambda SC & - & (\mu+\kappa)C \end{array}\right. .$$

Thus, it is sufficient to study this two-dimensional system in (S,C), and then use the relations

$$E=a-C$$
$$P=b-S-E$$

to get E and P. ◇

2.2 Preliminaries

Convention *In lists of function variables, "x" is short for "$x_1, x_2, \dots, x_n$".*

Definition 2.2.1. *Let $\dot{x}=F(x)$, where $F:\mathbb{R}^n\to\mathbb{R}^n$, be a continuous dynamical system.*

We say that $\dot{x}=F(x)$ is an n-dimensional system.

Assume that we seek functions defined on an interval $I \subset \mathbb{R}$ which obey the dynamics of the system. Then we say that I is the time set of the system.

Let $x : I \to \mathbb{R}^n$ be a function obeying the dynamics of the system, i.e. $\dot{x}(t) = F(x(t))$ for all $t \in I$. Then we say that x is a trajectory of the system, and that

$$x(I) = \{y \in \mathbb{R}^n \mid \exists t \in I : y = x(t)\}$$

is an orbit of the system. The (unique) orbit of which x_0 is an element is denoted $x(I, x_0)$.

We say that $\mathbb{R}^n$ is the state space of the system. The elements of the state space are called states.

Convention *In this thesis, all dynamical systems are continuous. Therefore, from now on, if nothing else is said, "dynamical system" will mean "continuous dynamical system".*

Let us also make the following convention.

Convention *When we say that $x : I \to \mathbb{R}^n$ obeys the dynamics of the system $\dot{x} = F(x)$ on I, we mean that $\dot{x}(t) = F(x(t))$ for all $t \in I$.*

Definition 2.2.2. *Let $\dot{x} = P(x)$, where $P(x) = (p_1(x), p_2(x), \ldots, p_n(x))$, with $p_i \in \mathbb{R}[x_1, x_2, \ldots, x_n]$. Then we say that $\dot{x} = P(x)$ is a polynomial dynamical system.*

2.3 Conservation laws

The algebraic relation $E(t) + C(t) \equiv \alpha$ in Example 2.1.1 is an example of a *conservation law.* In general, a conservation law of a dynamical system states that some function of the variables is constant under the dynamics of the system. In this thesis, however, we will limit ourselves to study a certain type of conservation law. Therefore, for convenience, we define the notion of a conseration law in this more limited sense.

Convention *The elements of vector spaces will be written as column vectors.*

Convention *When we write*

$$\begin{pmatrix} x_1 & x_2 & \ldots & x_n \end{pmatrix},$$

we mean the row vector with components x_1, x_2, $\ldots$, x_n. When we write

$$(x_1, x_2, \ldots, x_n),$$

we mean the n-tuple with x_1, x_2, $\ldots$, x_n as components. Each n-tuple corresponds to a column vector, by the convention above.

Definition 2.3.1. *Let $\dot{x} = F(x)$ be an n-dimensional dynamical system with time set I. Let $x : I \to \mathbb{R}^n$ be a function which obeys the dynamics of the system on I. Assume that $\sum_{j=1}^{n} \gamma_{ij} x_j(t)$ is constant on I. This fact is called a conservation law of the system, and we say that $\gamma_i = (\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{in}) \in \mathbb{R}^n$ defines a conservation law.*

Assume that $\gamma_i = (\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{in})$ defines a conservation law of a polynomial dynamical system $\dot{x} = F(x)$ and that the function $x : I \to \mathbb{R}^n$ obeys the dynamics of the system on I. Then $\sum_{j=1}^{n} \gamma_{ij} x_j(t)$ is constant. The constant to which this expression is equal can be determined by evaluating the expression in any point t; a natural choice is $t = 0$. Thus,

$$\sum_{j=1}^{n} \gamma_{ij} x_j(t) \equiv \sum_{j=1}^{n} \gamma_{ij} x_j(0).$$

Definition 2.3.2. *Assume that $\gamma_1, \gamma_2, \ldots, \gamma_k \in \mathbb{R}^n$ define conservation laws of a polynomial dynamical system. Let $\gamma_i^T = \begin{pmatrix} \gamma_{i1} & \gamma_{i2} & \ldots & \gamma_{in} \end{pmatrix}$ and let*

$$\Gamma = (\gamma_{ij})_{\substack{1 \le i \le k \\ 1 \le j \le n}} .$$

Then we say that Γ is the matrix corresponding to the conservation laws defined by $\gamma_i = (\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{in})\,, i = 1, 2, \ldots, k$.

Let $x_0 = x(0)$ (where $x(t)$ is still a function which obeys the dynamics of the system), let $\alpha_i(x_0) = \sum_{j=1}^{n} \gamma_{ij} x_j(0)$ and let

$$\alpha(x_0) = \begin{pmatrix} \alpha_1(x_0) & \alpha_2(x_0) & \ldots & \alpha_k(x_0) \end{pmatrix}^T .$$

Then $\Gamma x(t) = \alpha(x_0)$ for all t. This implies that $x(I, x_0)$ is a subset of the solution space of $\Gamma x = \alpha(x_0)$.

Definition 2.3.3. *Let* $\operatorname{sol}(A, b) = \{x \mid Ax = b\}$. *Then* $\operatorname{sol}(A, b)$ *is called the solution space of $Ax = b$.*

For $b = 0$, we write $\ker A$ *instead of* $\operatorname{sol}(A, 0)$.

Recall from linear algebra that the general solution of a non-homogeneous linear system of equations (i.e. $Ax = b$ where $b \neq 0$) is $x = x_h + x_p$, where x_h is a solution of the corresponding homogeneous equation (i.e. $Ax = 0$) and x_p is a solution of the non-homogeneous equation. In our case, the right-hand side is $\alpha(x_0)$, i.e. it depends on x_0. Thus,

$$\begin{aligned} \operatorname{sol}(\Gamma, \alpha(x_0)) &= \{x_p + x_h \mid x_p \in \operatorname{sol}(\Gamma, \alpha(x_0)) \text{ and } x_h \in \ker \Gamma\} \\ &= x_p(x_0) + \ker \Gamma, \end{aligned}$$

where $x_p(x_0)$ is a solution of $\Gamma x = \alpha(x_0)$. This is an *affine subspace* of $\mathbb{R}^n$; let us recall the definition of an affine subspace and the definition of the dimension of such a space.

Definition 2.3.4. *Let V be a vector space and let $A \subset V$ be a set such that $A = v + L = \{v + w \mid w \in L\}$ for some $v \in V$ and L a linear subspace of V. Then we say that A is an affine subspace of V.*

Let $A = v + L$ be a affine subspace of V. Then the dimension of the A is defined as $\dim L$.

Thus, the set of conservation laws defined by $\gamma_1, \gamma_2, \ldots, \gamma_k$ corresponds to the family of affine subspaces $\{x_p(x_0) + \ker \Gamma\}_{x_0 \in \mathbb{R}^n} \subset \mathbb{R}^n$. By picking an initial state, we pick one of these affine subspaces.

This is as good a time as any to recall from elementary linear algebra the well-known *rank-nullity-theorem*, which will be used several times throughout the thesis.

Proposition 2.3.5 ([8, Theorem 4.4])**.** *Let V and W be vector spaces over a field k and let $T : V \to W$ be a linear transformation. Then* $\dim V = \operatorname{rank}(T) + \dim\ker T$*.*

We can also make a geometrical interpretation of conservation laws. Assume that $(\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{in}) \in \mathbb{R}^n$ defines a conservation laws of a system. Then

$$\begin{pmatrix}\gamma_{i1} & \gamma_{i2} & \cdots & \gamma_{in}\end{pmatrix}\begin{pmatrix}x_1\\ x_2\\ \vdots\\ x_n\end{pmatrix} = \alpha_i$$

for some $\alpha_i \in \mathbb{R}$. Let $\gamma_i^T = \begin{pmatrix}\gamma_{i1} & \gamma_{i2} & \cdots & \gamma_{in}\end{pmatrix}$. Since γ_i^T can be interpreted as the matrix of a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}$, we can use Proposition 2.3.5, which gives that

$$n = \dim\ker\gamma_i^T + \operatorname{rank}\left(\gamma_i^T\right).$$

Now, $\operatorname{rank}\left(\gamma_i^T\right) = 1$ (since we can assume that not all γ_i equals zero), so $\dim\ker\gamma_i^T = n - 1$. Recall the definition of a *hyperplane*: a subset H of a vector space V is a hyperplane if and only if it is an affine subspace of dimension $n-1$. Thus,

$$H_{\gamma_i,x_0} = \left\{x \in \mathbb{R}^n \mid \sum_{j=1}^{n}\gamma_{ij}x_j = \sum_{j=1}^{n}\gamma_{ij}x_j(0)\right\}$$

is a hyperplane, and the conservation law defined by γ_i corresponds to the family of hyperplanes $\{H_{\gamma_i,x_0}\}_{x_0\in\mathbb{R}^n}$. A change in x_0 corresponds to a translation of the hyperplane. Now consider a set of conservation laws, each defined by $\gamma_i = (\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{in})$, for $i = 1, 2, \ldots, k$. For fixed $x_0 \in \mathbb{R}^n$, we have

$$x(I, x_0) \subset H(\gamma_i, x_0) \text{ for } i = 1, 2, \ldots, k,$$

so

$$x(I, x_0) \subset \bigcap_{i=1}^{k} H_{\gamma_i,x_0}.$$

An intersection of a set of hyperplanes is a *polyhedral set.* Thus, the orbit which x_0 belongs to is confined to a polyhedral set. Each hyperplane in the intersection which define the polyhedral set corresponds to a conservation law and x_0. A change in x_0 will translate each hyperplane, so a change in x_0 corresponds to a translation of the polyhedral set.

Let us return to the algebraic viewpoint. If $\gamma_1, \gamma_2, \ldots, \gamma_k \in \mathbb{R}^n$ define conservation laws of the system, we might ask ourselves if a proper subset of $\{\gamma_1, \gamma_2, \ldots, \gamma_k\}$ is enough to convey the same information about the system, i.e. whether some of the conservation laws are redundant. More precisely, there is redundancy if the solution space of $\Gamma x = \alpha$ is the same as the solution space of $\hat{\Gamma}x = \hat{\alpha}$, where $\hat{\Gamma}$ is Γ with some rows removed, and $\hat{\alpha}$ is α with the same rows removed, since then the conservation laws corresponding to the removed rows does not contribute any information which is not already conveyed by the conservation laws corresponding to the rows of $\hat{\Gamma}$. Now, by Definition 2.3.4, the dimension of the solution space of $\Gamma x = \alpha(x_0)$ equals $\dim\ker\Gamma$. This means that there is redundancy in a set of conservation laws if and only if $\dim\ker\Gamma = \dim\ker\hat{\Gamma}$. Since $n = \dim\ker\Gamma + \operatorname{rank}(\Gamma)$ and $n = \dim\ker\hat{\Gamma} + \operatorname{rank}\left(\hat{\Gamma}\right)$, we have

$$\dim\ker\Gamma - \dim\ker\hat{\Gamma} = \operatorname{rank}\left(\hat{\Gamma}\right) - \operatorname{rank}(\Gamma).$$

Thus, a conservation law is redundant if and only if $\operatorname{rank}\left(\hat{\Gamma}\right) = \operatorname{rank}(\Gamma)$, i.e. if and only if Γ does not have full rank; in other words, if and only if the γ_i are linearly independent in $\mathbb{R}^n$. Let us introduce the following terminology.

Definition 2.3.6. *Let $\gamma_i = (\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{in})$, $i = 1, 2, \ldots, k$, define conservation laws of a polynomial dynamical system. If $\{\gamma_i \mid i = 1, 2, \ldots, k\} \subset \mathbb{R}^n$ is linearly independent, we say that the conservation laws are linearly independent. Otherwise, we say that the conservation laws are linearly dependent.*

To summarize, a set of conservation laws gives us, for each initial state x_0, a superset of the orbit which x_0 belongs to. More precisely, the superset is $\operatorname{sol}(\Gamma, \alpha(x_0))$. Since

$$\dim\operatorname{sol}(\Gamma, \alpha(x_0)) = \dim\ker\Gamma,$$

this means that $\operatorname{sol}(\Gamma, \alpha(x_0))$ is a proper subset of $\mathbb{R}^n$ if and only if $\dim\ker\Gamma \neq n$, i.e. if and only if $\operatorname{rank}(\Gamma) \neq 0$. But if there is any conservation law at all, then $\operatorname{rank}(\Gamma) \geq 1$. Thus, the existence of a conservation law implies that $\operatorname{sol}(\Gamma, \alpha(x_0))$ is a proper subset of $\mathbb{R}^n$. Of course, this is as expected: the existence of a conservation law means precisely that the orbit cannot escape the corresponding hyperplane.

2.4 Example of algebraic technique for finding conservation laws

Later in this chapter, we will present an algebraic method for finding conservation laws of polynomial dynamical systems. The method will be based on the following example.

Example 2.4.1. This is a summary of [17, chapter 2.7.2].

In that chapter, networks of chemical reactions are studied. As in Example 2.1.1, it is assumed that law of mass action holds, which results in each variable x_i being governed by an equation of the form

$$\dot{x}_i = p_i(x_1, x_2, \ldots, x_n),$$

where $p_i \in \mathbb{R}[x_1, x_2, \ldots, x_n]$, so the network is governed by a polynomial dynamical system. It is then noted that by gathering each monomial which appears in any of the p_i and putting them, in some order, in a column vector m, we can write the system on the form

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \vdots \\ \dot{x}_n \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1k} \\ c_{21} & c_{22} & \cdots & c_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nk} \end{pmatrix} m$$

for some $c_{ij} \in \mathbb{R}$. The matrix

$$C = (c_{ij})_{\substack{1 \le i \le n \\ 1 \le j \le k}}$$

(called Γ in [17]) is called a stoichiometry matrix. Let

$$\dot{x} = \begin{pmatrix} \dot{x}_1 & \dot{x}_2 & \dots \dot{x}_n \end{pmatrix}^T .$$

Then $\dot{x} = Cm$. Let

$$v = \begin{pmatrix} v_1 & v_2 & \dots & v_n \end{pmatrix} .$$

Then, it is noted, if $v\Gamma = 0$ (i.e. if v is in the left kernel of Γ), then $\sum_{i=1}^k v_i \dot{x}_i = 0$. By integration of both sides of the equation, we get $\sum_{i=1}^k v_i x_i(t) \equiv \alpha$, for some $\alpha \in \mathbb{R}$, which is a conservation law. Finally, it is remarked that the number of linearly independent conservation laws of the system is given by the dimension of the left kernel of Γ. ◇

Example 2.4.1 shows how to find conservation laws of stoichiometry systems. But note that the only thing which was used was that the dynamical system had a polynomial right-hand side. Thus, this method can be generalized to all polynomial dynamical systems.

In the rest of this section, we will make this method for finding conservation laws of polynomial dynamical systems precise.

2.5 Matrix representation of a polynomial dynamical system

Convention *In lists of variables of a polynomial ring, "x" is short for "x_1, x_2, ..., x_n".*

Convention *k denotes a field.*

Convention *In this thesis,* $\mathbb{N} = \{0, 1, 2, \dots\}$.

Definition 2.5.1. *Let $m = \prod_{i=1}^n x_i^{\alpha_i} \subset k[x]$ for some $\alpha_i \in \mathbb{N}$. Then we say that m is a monomial.*

Convention *The monomial $\prod_{i=1}^n x_i^0 \in k[x]$ is denoted* 1. *This element is not to be confused with the multiplicative identity element of k.*

Consider a polynomial $p \in k[x]$. A polynomial is a linear combination of monomials. It is clear that, if we do not allow zeros as coefficients, the representation of a polynomial in terms of monomials is unique (up to reordering of the monomials). Thus, the following definition makes sense.

Definition 2.5.2 ([7, chapter 3.2.2])**.** *Let $f = \sum_{i=1}^r c_i m_i \in k[x]$, where $c_i \neq 0$ for all i. Then we say that* $\operatorname{supp}(f) = \{m_1, m_2, \dots, m_r\}$ *is the support of f.*

For convenience, let us generalize this a bit.

Definition 2.5.3 ([7, chapter 3.2.2])**.** *Let $P \subset k[x]$ be a set. We say that* $\operatorname{supp}(P) = \cup_{p \in P} \operatorname{supp}(p)$ *is the support of P.*

Let $P = \{p_1, p_2, \dots, p_k\}$ be a set of polynomials. If we allow zeros as coefficients, we can write each $p_i \in P$ as $p_i = \sum_{m \in \operatorname{supp}(P)} c_{i,m} m$, for some $c_{im} \in \mathbb{R}$; in other words,

$$p_i = \begin{pmatrix} c_{i,m_1} & c_{i,m_2} & \dots & c_{i,m_p} \end{pmatrix} \begin{pmatrix} m_1 \\ m_2 \\ \dots \\ m_p \end{pmatrix},$$

where $\operatorname{supp}(P) = \{m_1, m_2, \dots, m_p\}$. Let

$$\begin{aligned} p &= \begin{pmatrix} p_1 & p_2 & \dots & p_k \end{pmatrix}^T, \\ m &= \begin{pmatrix} m_1 & m_2 & \dots m_p \end{pmatrix}^T, \text{ and} \\ C &= (c_{i,m_j})_{\substack{1 \le i \le k \\ 1 \le j \le p}}. \end{aligned}$$

Then $p = Cm$.

Example 2.5.4. Let $p_i \in k[x, y]$, $i = 1, 2, 3, 4$, where

$$\begin{aligned} p_1 &= x^2y + xy^2, \\ p_2 &= x^2 + xy^2 - y, \\ p_3 &= xy^2 + y + 3, \text{ and} \\ p_4 &= y^2. \end{aligned}$$

Let $P = \{p_i \mid i \in \{1, 2, 3, 4\}\}$. Then $\operatorname{supp}(P) = \{x^2y, x^2, xy^2, y^2, y, 1\}$. Then we can write

$$\begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 3 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} x^2y \\ x^2 \\ xy^2 \\ y^2 \\ y \\ 1 \end{pmatrix}.$$

But we can also write for example

$$\begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 3 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} xy^2 \\ y^2 \\ x^2y \\ y \\ x^2 \\ 1 \end{pmatrix},$$

or

$$\begin{pmatrix} p_3 \\ p_1 \\ p_4 \\ p_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 3 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & 1 & 0 \end{pmatrix} \begin{pmatrix} xy^2 \\ y^2 \\ x^2y \\ y \\ x^2 \\ 1 \end{pmatrix}.$$

The three expressions above all have the form $p = Cm$, where

$$p = \left(p_{\sigma(1)}, p_{\sigma(2)}, p_{\sigma(3)}, p_{\sigma(4)}\right)$$

for some permutation σ of $\{1,2,3,4\}$ and m is a column vector with the elements of $\operatorname{supp}(P)$, in some order, as components, but the matrix C depends on the order of the components of p and order of the components of m. $\diamond$

Convention *Often — as in the example above — when speaking of a polynomial $p \in k[x_1, x_2, \ldots, x_n]$ we will write p_i instead of $p_i(x_1, x_2, \ldots, x_n)$ or $p_i(x)$, since the variables of p will always be clear from context. However, when a polynomial is the right-hand side of a differential equation, we write $\dot{x} = p(x)$, not $\dot{x} = p$.*

Example 2.5.4 shows that the matrix C, defined before the example, is not unique for a set of polynomials. We want to be able to speak of *the* matrix representation of a set of polynomials. Let us turn to this problem.

Let us introduce some temporary terminology. Given a set of polynomials P, let us call

- an expression of the form $p = Cm$ (where p, C and m are defined as above) a matrix representation of P,
- the vector p a vector of the polynomials in P,
- the matrix C a coefficient matrix, corresponding to the order of the components of p and m, of P, and
- the vector m a vector of the monomials in $\operatorname{supp}(P)$.

Convention *If S is a set, then $|S|$ denotes the number of elements of S. If S is infinite, then $|S| = \infty$.*

Let $k = |P|$ and $s = |\operatorname{supp}(P)|$. Since there are k vectors consisting of the polynomials in P and s vectors consisting of the monomials in $\operatorname{supp}(P)$, there are $k \cdot s$ coefficient matrices of P. Given an order of the elements of P and $\operatorname{supp}(P)$, however, there is a unique coefficient matrix.

Let us define a concept which is convenient to use for talking about the order in which the elements of P are listed in the vector p.

Definition 2.5.5. *Let $P \subset k[x]$ be a finite set with $|P| = k$. Let*

$$\mu : \{1, 2, \ldots, k\} \to P$$

be a bijection. Then μ is called an enumeration of P.

Let μ be an enumeration of P, with $|P| = k$. The notation

$$p_\mu = \begin{pmatrix} \mu(1) & \mu(2) & \ldots & \mu(k) \end{pmatrix}^T$$

will be used.

A choice of enumeration of P fixes the order of the rows of the coefficient matrix of P.

Next, we want to introduce terminology for talking about the order in which the elements of $\operatorname{supp}(P)$ are listed in the vector m. For the purpose of matrix representations, we could just make an arbitrary choice of an order in which to list the elements. However, the notion of *monomial orderings* is an established concept in the literature, and it will be important later. Therefore, we will require that the order in which the monomials are listed in the vector of monomials in $\operatorname{supp}(P)$ satisfy some monomial ordering. This makes some orders in which to list the monomials in $\operatorname{supp}(P)$ inadmissible, but for our purposes, this is no loss. Also, it enables us to use one concept for multiple purposes.

A priori, neither the monomials in one variable, nor the monomials in $n > 1$ variables, are ordered. However, for monomials in one variable, we often implicitly order them by their degree, which is of course very natural. It is even the only possible criterion by which to order monomials, since the degree is the only thing distinguishing one monomial from another. For monomials in $n > 1$ variables, on the other hand, many different ways to order the monomials are conceivable. This leads us the the notion of monomial orderings. First, recall the definition of an order on a set.

Definition 2.5.6 ([15, Definition 1.5])**.** *An order $<$ on a set S is a relation such that*

- *for every pair of elements $x, y \in S$, precisely one of the following statements holds:*
$$x < y,\ x = y,\ y < x,$$
and
- *if $x < y$, then $x + z < y + z$ for any $z \in S$.*

Remark *To distingush this from a partial order on a set, this concept is sometimes called a "total order on a set".*

Now we can define the notion of a monomial ordering.

Convention $\operatorname{mon}(k[x]) = \{m \in k[x] \mid m \textit{ monomial }\}$.

Definition 2.5.7. *A monomial ordering on $k[x]$ is an order $<$ on $\operatorname{mon}(k[x])$ such that*

- *$1 < m$ for all $m \in \operatorname{mon}(k[x])$, and*
- *if $m_1 < m_2$, then $mm_1 < mm_2$ for every $m \in \operatorname{mon}(k[x])$*

[7, chapter 3.1].

Let $<$ be a monomial ordering on $k[x]$. Let $\operatorname{supp}(P) = \{m_1, m_2, \ldots, m_s\}$, where $m_i > m_{i+1}$ for all i. We define

$$m_< = \begin{pmatrix} m_1 & m_2 & \ldots & m_s \end{pmatrix}^T.$$

Convention *$x > y$ if and only if $y < x$.*

A choice of a monomial ordering fixes the order of the columns of the coefficient matrix of C.

Now we are ready to make the notions of a matrix representation and a coefficient matrix of a set of polynomials permanent.

Definition 2.5.8. *Let $P \subset k[x]$ be a finite set. Let μ be an enumeration of P and let $<$ be a monomial ordering on $k[x]$. Let $C_{\mu,<}$ be the unique matrix which satisfies $p_\mu = C_{<,\mu} m_<$. Then*

$$p_\mu = C_{<,\mu} m_<$$

is called the matrix representation, and $C_{<,\mu}$ is called the coefficient matrix, of P corresponding to the enumeration μ and the monomial ordering $<$.

So far, we have talked about sets of polynomials. Let us now turn to what this means for polynomial dynamical systems. Let $\dot{x} = F(x)$ be an n-dimensional polynomial dynamical system, i.e. $F(x) = (p_1(x), p_2(x), \ldots, p_n(x))$ for some $p_i \in k[x]$. Let $P = \{p_1, p_2, \ldots, p_n\}$. Let μ be any enumeration of P (so it is possible that μ is an enumeration such that $\mu(i) \neq p_i$) and let $<$ be a monomial ordering of $k[x]$. Then $p_\mu = C_{<,\mu} m_<$. Let

$$x_\mu = \begin{pmatrix} x_{\mu^{-1}(p_1)} & x_{\mu^{-1}(p_2)} & \cdots & x_{\mu^{-1}(p_n)} \end{pmatrix}^T .$$

Then $\dot{x}_\mu = C_{<,\mu} m_<$. However, we usually already have an implicit ordering of the variables $x_1, x_2, \ldots, x_n$, and $\dot{x}_i = p_i(x)$, so the natural enumeration of P is to let μ be the identity on $\{1, 2, \ldots, n\}$. Let us make this the convention for this thesis. This leads us to the following definition.

Definition 2.5.9. *Let $\dot{x} = F(x) = (p_1(x), p_2(x), \ldots, p_n(x))$ be an n-dimensional polynomial dynamical system. Let*

- *$P = \{p_1, p_2, \ldots, p_n\}$,*
- *$\mu = \mathrm{id}_{\{1,2,\ldots,n\}}$, where $\mathrm{id}_{\{1,2,\ldots,n\}}$ denotes the identity function of $\{1, 2, \ldots, n\}$,*
- *$<$ be a monomial ordering on $\mathbb{R}[x]$,*
- *$p_\mu = C_{<,\mu} m_<$ be the matrix representation of P corresponding to μ and $<$, and*
- *$C_< = C_{<,\mu}$.*

Then $\dot{x} = C_< m_<$ is called the matrix representation, and $C_<$ is called the coefficient matrix, of $\dot{x} = F(x)$ corresponding to $<$.

We will work with two classes of monomial orderings: *Lex-orderings* (”Lex” stands for ”lexicographic”) and *Deglex-orderings* (”Deglex” stands for ”degree lexicographic”).

Definition 2.5.10 ([7, chapter 3.1])**.** *Let $\sigma \in S_n$ be a permutation of $\{1, 2, \ldots, n\}$. Let $\sigma(j) = i_j$ for $j = 1, 2, \ldots, n$. The Lex-ordering corresponding to σ is the ordering $<$ satisfying that*

$$\prod_{j=1}^{n} x_j^{\alpha_j} < \prod_{j=1}^{n} x_j^{\beta_j}$$

if and only if there is some $k \in \{1, 2, \ldots, n\}$ such that

- $\alpha_{i_j} = \beta_{i_j}$ *for* $1 \leq j < k$ *and*
- $\alpha_{i_k} < \beta_{i_k}$.

Then we write

$$\prod_{j=1}^{n} x_j^{\alpha_j} <_{\text{Lex}(\sigma)} \prod_{j=1}^{n} x_j^{\beta_j}.$$

After we have defined Deglex-orderings, we will give examples illustrating both Lex- and Deglex-orderings. Before defining Deglex, however, we must define the notion of degree of monomials and polynomials in n variables. A monomial in the polynomial ring in one variable is simple: it is just the variable to some power, and the power is called the degree of the monomial. The degree of a polynomial in one variable is the degree of the monomial with maximum degree. For monomials in $n \geq 1$ variables, we need two distinct but related concepts.

Definition 2.5.11 ([4, Definition 7 in chapter 2])**.** *Let* $m = \prod_{i=1}^{n} x_i^{\alpha_i} \in k[x]$.

- *The multidegree of* m *is defined as the* n*-tuple* $(\alpha_1, \alpha_2, \ldots, \alpha_n)$*, and*
- *the degree of* m *is defined as* $\sum_{i=1}^{n} \alpha_i$.

A monomial in one variable is characterized by its degree: there is only one monomial for every degree. This is not true for monomials in $n > 1$ variables. E.g. x^2 and xy in $k[x, y]$ both have degree two, but they are not the same. Instead, a monomial is characterized by its multidegree: there is a one-to-one correspondence between n-tuples of natural numbers and monomials in the polynomial ring in n variables. Now we can define the degree of a polynomial in several variables.

Definition 2.5.12 ([4, Definition 1 and 3 in chapter 1])**.** *Let* $f \in k[x]$. *Then the degree of* f *is defined as* $\max\{\deg(m) \mid m \in \operatorname{supp}(f)\}$.

Example 2.5.13. Let

$$f = x_1^2 x_2 + x_2 x_3 + x_1 x_2^2 x_3.$$

Let

$$\begin{aligned} m_1 &= x_1^2 x_2, \\ m_2 &= m_2 = x_2 x_3, \text{ and} \\ m_3 &= x_1 x_2^2 x_3. \end{aligned}$$

Then $\operatorname{supp}(f) = \{m_1, m_2, m_3\}$. Since

$$\begin{aligned} \deg(m_1) &= 2 + 1 = 3, \\ \deg(m_2) &= 1 + 1 = 2, \text{ and} \\ \deg(m_3) &= 1 + 2 + 1 = 4 \end{aligned}$$

we get $\deg(f) = \max\{2, 3, 4\} = 4$. ◇

Now we can introduce Deglex.

Definition 2.5.14 ([7, chapter 3.1])**.** *Let σ be as in Definition 2.5.10. The Deglex-ordering corresponding to σ is the ordering $<$ satisfying that*

$$\prod_{j=1}^{n} x_j^{\alpha_j} < \prod_{j=1}^{n} x_j^{\beta_j}$$

if and only if either

(i) $\sum_{j=1}^{n} \alpha_j < \sum_{j=1}^{n} \beta_j$*, or*

(ii) $\sum_{j=1}^{n} \alpha_j = \sum_{j=1}^{n} \beta_j$ *and* $\prod_{j=1}^{n} x_j^{\alpha_j} <_{\mathrm{Lex}(\sigma)} \prod_{j=1}^{n} x_j^{\beta_j}$.

Then we write

$$\prod_{j=1}^{n} x_j^{\alpha_j} <_{\mathrm{Deglex}(\sigma)} \prod_{j=1}^{n} x_j^{\beta_j}.$$

Example 2.5.15. To illustrate Lex- and Deglex-orderings, let us consider some monomials in $k[x_1, x_2]$.

First, let σ be the identity permutation of $\{1, 2\}$, i.e. $\sigma(j) = j$ for $j \in \{1, 2\}$. This corresponds to the Lex-ordering with $x_1 > x_2$. To see this, note that $x_1 = x_1^1 x_2^0$ and $x_2 = x_1^0 x_2^1$. In other words,

$$\begin{aligned} \alpha_1 &= 1, \quad \alpha_2 = 0, \\ \beta_1 &= 0, \quad \beta_2 = 1. \end{aligned}$$

in the notation of the definition. The permutation is trivial, i.e. $i_1 = \sigma(1) = 1$ and $i_2 = \sigma(2) = 2$. Thus, we shall first compare α_1 with β_1. We see that $\alpha_1 > \beta_1$, so we can take $k = 1$, where k is as in the definition. Thus, $x_1 > x_2$.

Consider x_1 and x_2^m, where $m > 1$. Since, again, x_1 has more x_1-factors than x_2^m has, we have $x_2 <_{\mathrm{Lex}(\sigma)} x_1$, but since x_2^m has higher degree than x_1 has, we have $x_1 <_{\mathrm{Deglex}(\sigma)} x_2$.

Now let σ be the permutation of $\{1, 2\}$ with $\sigma(1) = 2$ and $\sigma(2) = 1$. Then $x_1 <_{\mathrm{Lex}(\sigma)} x_2^m$ for every m, but $x_2 <_{\mathrm{Deglex}(\sigma)} x_1^m$ for every $m > 1$. ◇

We will usually not define σ formally: instead we will speak of, for example, "the Lex-ordering with $x_2 > x_3 > x_1$", which corresponds to the permutation σ of $\{1, 2, 3\}$ with $\sigma(1) = 2$, $\sigma(2) = 3$ and $\sigma(3) = 1$.

Let us illustrate the concept of matrix representations of polynomial dynamical systems by an example.

Example 2.5.16. Consider the system $\dot{x}_i = p_i(x),\ i = 1, 2, 3, 4$, where

$$\begin{aligned} p_1 &= x_1^2 x_2 x_4 + x_1 x_2^2 x_4 + 2x_3 x_4 - x_4 \\ p_2 &= 2x_1^2 x_2^2 x_3 - 3x_1 x_2^2 x_4 + x_3 x_4 \\ p_3 &= 2x_1^2 x_2 x_4 - 8x_1^2 x_2^2 x_3 + 14x_1 x_2^2 x_4 - 2x_4 \\ p_4 &= 3x_1^2 x_2 x_4 + 2x_1^2 x_2^2 x_3 + 7x_3 x_4 - 3x_4 \end{aligned}$$

Let $<$ be the Lex-ordering with $x_1 > x_2 > x_3 > x_4$. Then

$$m_<^T = \begin{pmatrix} x_1^2 x_2^2 x_3 & x_1^2 x_2 x_4 & x_1 x_2^2 x_4 & x_3 x_4 & x_4 \end{pmatrix}$$

and

$$C_< = \begin{pmatrix} 0 & 1 & 1 & 2 & -1 \\ 2 & 0 & -3 & 1 & 0 \\ -8 & 2 & 14 & 0 & -2 \\ 2 & 3 & 0 & 7 & -3 \end{pmatrix}$$

◇

2.6 Matrix representation and conservation laws

Definition 2.6.1 ([9, Chapter 1.§3.2]). *Let A be square matrix. If $\det A \neq 0$, we say that A is non-singular.*

Let $<_1$ and $<_2$ be two different monomial orderings. Then $C_{<_1}$ and $C_{<_2}$ only differ in the order of the columns. Assume that $C_{<_1}$ and $C_{<_2}$ be $n \times m$-matrices. Let

$$E_{ij} = (e_{ijrs})_{\substack{1 \leq r \leq n \\ 1 \leq s \leq n}}$$

with

$$e_{ijrs} = \begin{cases} 1, & (r \neq i \text{ and } s \neq j) \text{ or } (r = j \text{ and } s = i) \text{ or } (r = i \text{ and } s = j) \\ 0, & \text{otherwise} \end{cases}.$$

It is clear that $\det E_{ij} = -1$. It is also clear that AE_{ij} is the matrix A with columns i and j switched. Since $C_{<_1}$ and $C_{<_2}$ differ only in the order of the columns, there are tuples $\{(i_1, j_1), (i_2, j_2), \ldots, (i_k, j_k)\}$ such that $C_{<_1} = C_{<_2} \prod_{r=1}^{k} E_{i_r j_r}$. Let $E = \prod_{r=1}^{k} E_{i_r j_r}$. Since $\det E^T = \det E = (-1)^r \neq 0$, Proposition A.0.1 implies that

$$\ker (C_{<_1})^T = \ker \left(E^T C_{<_2}^T\right) = \ker (C_{<_2})^T,$$

since E^T is non-singular.

Proposition 2.6.2. *Let $\dot{x}_i = p_i(x)$, $i = 1, 2, \ldots, n$, be a polynomial dynamical system. Let $\dot{x} = C_< m_<$ be the matrix representation of this system corresponding to monomial ordering $<$. Then*

$$\gamma^T = \begin{pmatrix} \gamma_1 & \gamma_2 & \ldots & \gamma_n \end{pmatrix} \in \mathbb{R}^n$$

defines a conservation law if and only if $\gamma \in \ker C_<^T$.

Proof. Assume that $\gamma^T = (\gamma_1, \gamma_2, \ldots, \gamma_n)$ defines a conservation law, i.e.

$$\sum_{i=1}^{n} \gamma_i x_i \equiv \alpha$$

for some $\alpha \in \mathbb{R}$, for all functions x which obey the dynamics of the system. Taking the derivative of both sides gives $\sum_{i=1}^{n} \gamma_i \dot{x}_i = 0$, i.e.

$$0 = \gamma^T \dot{x} = \gamma^T C_< m_<,$$

so γ^T is in the left kernel of $C_<$, which is equivalent to $c \in \ker (C_<)^T$.

On the other hand, assume that $\gamma \in \ker (C_<)^T$. Then $\gamma^T (C_<) = 0$, so

$$0 = \gamma^T (C_<) m = \gamma^T \dot{x}.$$

By integrating both sides we get $\alpha = \gamma^T x$, i.e. a conservation law. □

This leads us to formulate the following proposition.

Proposition 2.6.3. *Let $\dot{x}_i = p_i(x)$, $i = 1, 2, \ldots, n$, be a polynomial dynamical system. Let $P = \{p_1, p_2, \ldots, p_n\}$. Let $<$ be a monomial ordering on* mon $(\mathbb{R}[x])$. *Let $C_<$ be the coefficient matrix of P corresponding to $<$. Then $\dot{x} = p_i(x)$, $i = 1, 2, \ldots, n$, can be reduced to a* rank $(C_<)$*-dimensional system (i.e. the dynamics of the system can be expressed using $r = \operatorname{rank}(C_<)$ variables).*

Proof. Let $k = \dim\ker\left(C_<\right)^T$. Let $\{\gamma_1, \gamma_2, \ldots, \gamma_k\}$ be a basis of $\ker C_<^T$. In particular,

$$\gamma_i \in \ker C_<^T \subset \mathbb{R}^n$$

for $i = 1, 2, \ldots, k$. Thus, each $\gamma_i = \begin{pmatrix}\gamma_{i1} & \gamma_{i2} & \ldots & \gamma_{in}\end{pmatrix}^T$ defines a conservation law, namely

$$\sum_{j=1}^{n} \gamma_{ij} x_j(t) \equiv \alpha_j,$$

for some $\alpha_j \in \mathbb{R}$. Since $\{\gamma_1, \gamma_2, \ldots, \gamma_k\}$ is a basis, this means, in particular, that the conservation laws generated by the γ_i, $i = 1, 2, \ldots, k$, are linearly independent in $\mathbb{R}^n$. Let

$$\Gamma = \left(\gamma_{ij}\right)_{\substack{1\le i\le k\\ 1\le j\le n}}.$$

The rows of Γ are linearly independent, since the γ_i are linearly independent. This means

$$\operatorname{rank}\left(\Gamma\right) \geq k.$$

On the other hand, we know that

$$\operatorname{rank}\left(\Gamma\right) \leq \min\left\{k, n\right\} \leq k,$$

so $\operatorname{rank}\left(\Gamma\right) = k$. In other words, Γ has full rank. This implies that there is a non-singular matrix E and an aribtrary matrix A such that

$$E\Gamma = \begin{pmatrix}A & I_k\end{pmatrix}$$

(E is in fact the product of the so called *elementary matrices* corresponding to the appropriate elementary row operations). Let $\beta = E\alpha$. Thus, the system $\Gamma x = \alpha$ has the solution

$$\begin{pmatrix}x_{n-k+1}\\ x_{n-k+2}\\ \vdots\\ x_n\end{pmatrix} = \begin{pmatrix}\beta_1 - \sum_{j=1}^{n-k} a_{1j}x_j\\ \beta_2 - \sum_{j=1}^{n-k} a_{2j}x_j\\ \vdots\\ \beta_k - \sum_{j=1}^{n-k} a_{kj}x_j\end{pmatrix} = \beta - A\tilde{x}$$

where

$$A = \left(a_{ij}\right)_{\substack{1\le i\le k\\ 1\le j\le n-k}}$$

and $\tilde{x} = \begin{pmatrix}x_1 & x_2 & \ldots x_{n-k}\end{pmatrix}^T$. This is a solution with k parameters. Thus, an n-dimensional system can be reduced to a $(n-k)$-dimensional system. Finally, note that

$$\begin{aligned} n - k &= n - \dim\ker\left(C_<^T\right)\\ &= \operatorname{rank}\left(C_<^T\right)\\ &= \operatorname{rank}\left(C_<\right)\end{aligned}$$

by Proposition 2.3.5. □

As was mentioned above, we can find $k = \dim\ker {C_<}^T$ linearly independent conservation laws. If two conservation laws are lineraly independent, they each contribute new information about the system. By what has been said above, we know that we can find at least k conservation laws, each properly contributing information about the system. But we also know that there is no point in trying to find more than k conservation laws, since when we already have k linearly independent conservation laws, any new conservation law will already be implicit in those that we have already found.

Example 2.6.4. We continue Example 2.5.16. The rank of $C_<$ is 2. Hence, $\dim\ker C_<^T = 4 - 2 = 2$. The set

$$\left\{ \begin{pmatrix} -3 & -1 & 0 & 1 \end{pmatrix}^T, \begin{pmatrix} -2 & 4 & 1 & 0 \end{pmatrix}^T \right\}$$

is a basis for $\ker C_<^T$. This gives two conservation laws:

$$\begin{array}{rcrcrcrcl} -3x_1(t) & - & x_2(t) & & & + & x_4(t) & \equiv & \alpha_1 \\ -2x_1(t) & + & 4x_2(t) & + & x_3(t) & & & \equiv & \alpha_2 \end{array}$$

for some $\alpha_1, \alpha_2 \in \mathbb{R}$. Solving for x_3 and x_4 and substituting into $m_<$ gives $m_< = T\tilde{m}_<$, where

$$T = \begin{pmatrix} 0 & 3 & 0 & 1 & \alpha_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & -4 & \alpha_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 & 1 & \alpha_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 6 & 0 & 0 & -10 & \beta_1 & -4 & \beta_2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 1 & \alpha_1 \end{pmatrix}$$

where $\beta_1 = 2\alpha_1 + 3\alpha_2$ and $\beta_2 = \alpha_2 - 4\alpha_1$, and

$$\begin{aligned} m_<^T = (\; & x_1^3x_2^2 \quad x_1^3x_2 \quad x_1^2x_2^3 \quad x_1^2x_2^2 \quad x_1^2x_2 \quad x_1^2 \quad x_1x_2^3 \\ & x_1x_2^2 \quad x_1x_2 \quad x_1 \quad x_2^2 \quad x_2 \quad 1 \;). \end{aligned}$$

Thus,

$$\dot{x} = C_<m_<(x) = C_<T\tilde{m}_<(x).$$

Since x_3 and x_4 has been eliminated, we only need the first and second row, denoted $(C_<)_1$ and $(C_<)_2$, of $C_<$. Thus, the matrix representation of the reduced system is

$$\dot{\tilde{x}} = \tilde{C}_<\tilde{m}_<(x)$$

where $\tilde{x}^T = \begin{pmatrix} x_1 & x_2 & \dots & x_k \end{pmatrix}$ and

$$\begin{aligned} \tilde{C}_< &= \begin{pmatrix} (C_<)_1 \\ (C_<)_2 \end{pmatrix} T \\ &= \left(\begin{matrix} 0 & 3 & 0 & 4 & \alpha_1 & 12 & 1 & \alpha_1 & -20 \\ 4 & 0 & -8 & -9 + 2\alpha_2 & 0 & 6 & -3 & -3\alpha_1 & -10 \end{matrix} \right. \\ & \qquad \left. \begin{matrix} -3 + 2\beta_1 & -8 & -1 + 2\beta_2 & -\alpha_1 \\ \beta_1 & -4 & \beta_2 & 0 \end{matrix} \right). \end{aligned}$$

In other words, the reduced system is $\dot{x}_i = q_i(x),\ i = 1, 2$, where

$$\begin{aligned} q_1 &= 3x_1^3x_2 + 4x_1^2x_2^2 + \alpha_1 x_1^2 x_2 + 12x_1^2 + x_1x_2^3 + \alpha_1 x_1 x_2^2 - 20x_1x_2 \\ &\quad + (-3 + 2\beta_1)x_1 - 8x_2^2 + (-1 + 2\beta_2)x_2 - \alpha_1 \\ q_2 &= 4x_1^3x_2^2 - 8x_1^2x_2^3 + (-9 + 2\alpha_2)x_1^2x_2^2 + 6x_1^2 - 3x_1x_2^3 - 3\alpha_1 x_1 x_2^2 - 10x_1x_2 \\ &\quad + \beta_1 x_1 - 4x_2^2 + \beta_2 x_2. \end{aligned}$$

◇

3 Finding the steady states

3.1 Introduction

Definition 3.1.1. *We say that $\hat{x} \in \mathbb{R}^n$ is a steady state of the n-dimensional system $\dot{x} = F(x)$ if $F(\hat{x}) = 0$.*

Let $\dot{x}_i = p_i(x)$, $i = 1, 2, \ldots, n$, be a polynomial dynamical system. Then the steady states are the real solutions (i.e. those in $\mathbb{R}^n$) of the system of polynomial equations

$$p_i(x) = 0,\ i = 1, 2, \ldots, n.$$

Unless $\deg(p_i) \leq 1$ for all i, this system of equations is non-linear. Solving a non-linear system of equations is typically difficult. However, there are methods available for tackling such problems. We will present one such method, which is based on so called *Gröbner bases*. It will allow us to find the common roots of a set of multivarite polynomials by finding the roots of a sequence of univariate polynomials.

3.2 A remark on matrix respresentations and steady states

Let $<$ be a monomial ordering of $\mathbb{R}[x]$ and let $\dot{x} = C_< m_<$ be the corresponding matrix representation of a polynomial dynamical system $\dot{x}_i = p_i(x),\ i = 1, 2, \ldots, n$. In the previous chapter, we saw that we can use $\ker C_<^T$ to find the conservation laws of a system, and $\operatorname{im} C_<$ to find a linear subspace $S \subset \mathbb{R}^n$ such that each orbit of the system is a subset of a coset of S. Before we begin our presentation of the method for finding steady states, let us make a small remark about how $\ker C_<$ can be interpreted.

If $\hat{x} \in \mathbb{R}^n$ is such that

$$y = m_<(\hat{x}) \in \ker C_<,$$

then

$$C_< m_<(\hat{x}) = 0,$$

so $\hat{x}$ is a steady state. On the other hand, if $\hat{x}$ is a steady state, then

$$C_< m_<(\hat{x}) = 0,$$

so $\hat{x}$ is such that

$$m_<(\hat{x}) \in \ker C_<.$$

In other words, the set of steady states of the system is precisely the preimage of $\operatorname{im} m \cap \ker C_<$ under

$$\begin{array}{rccc} m: & \mathbb{R}^n & \to & \mathbb{R} \\ & x & \mapsto & m_<(x) \end{array}.$$

3.3 Ideals and varieties

Let $p_i \in k[x]$, $i = 1, 2, \ldots, n$. Assume that $\hat{x}$ is a simultaneous root of all the p_i, i.e. $p_i(\hat{x}) = 0$ for $i = 1, 2, \ldots, n$. Then

$$\sum_{i=1}^{n} p_i(\hat{x}) q_i(\hat{x}) = 0$$

for all $q_i \in k[x]$. On the other hand, consider the set

$$I = \left\{ \sum_{i=1}^{n} p_i q_i \mid q_i \in k[x] \right\} \subset k[x].$$

Let $\hat{x}$ be a simultaneous root of all polyonomials in I. Then, in particular, $p_i(\hat{x}) = 0$ for all i. Thus, $\hat{x}$ is a simultaneous root of $p_i, i = 1, 2, \ldots, n$ if and only if it is a root of every element in I.

The set I is called an *ideal* of $k[x]$, and the set of common roots of I is called a *variety*. More generally, ideals and varieties are defined as follows.

Definition 3.3.1 ([1, chapter 1]). *Let R be a commutative ring with identity. Let $I \subset R$ be a set such that*

- *$j_1 + j_2 \in I$ for all $j_1, j_2 \in I$, and*
- *$rj \in I$ for all $r \in R$ and $j \in J$.*

Then we say that I is an ideal.

Let $S \subset R$ be a set. Then

$$I = \langle S \rangle = \left\{ \sum_{j=1}^{k} r_i s_i \mid r_i \in R, s_i \in S, k \in \{1, 2, \ldots, n\} \right\}$$

is called the ideal generated by S.

If $S = \{s_1, s_2, \ldots, s_n\}$, we sometimes write $\langle s_1, s_2, \ldots, s_n \rangle$ for the ideal generated by S.

Convention *From now on, whenever we say "ring", we mean "commutative ring with identity".*

Before giving the definition of a variety, we need to recall the definition of the *algebraic closure* of a field.

Convention *$k[t]$ will denote the polynomial ring in precisely one variable.*

Definition 3.3.2 (see e.g. [7, end of chapter 1.3.3]). *Let k be a field. Let $\overline{k} \supset k$ be the smallest field with the following property: for each $p \in k[t]$, each root of p belongs to $\overline{k}$. Then $\overline{k}$ is called the algebraic closure of k.*

For example, $\overline{\mathbb{R}} = \mathbb{C}$, as is very well-known.

Definition 3.3.3 (e.g. [4]). *Let $I \subset k[x]$. The set*

$$V(I) = \left\{ \alpha \in \overline{k}^n \mid \forall p \in I : \ p(\alpha) = 0 \right\},$$

where $\overline{k}$ denotes the algebraic closure of k, is called the variety of I.

Remark *Another name for the notion called a variety is "zero set". However, we will use the term "variety", since it is the convention in commutative algebra.*

Definition 3.3.4 (cf. [12]). *Let $I \subset k[x]$. For fields K such that $k \subset K \subset \overline{k}$, we will use the notation*

$$V_K(I) = V(I) \cap K^n.$$

Let us illustrate the notion of a variety with a simple example.

Example 3.3.5. Let

$$I = \langle x^2 - y,\ x - y \rangle \subset k[x, y].$$

By the remarks above, $V(I)$ is given by the solutions of the system

$$\begin{cases} p_1 = x^2 - y = 0 \\ p_2 = x - y = 0 \end{cases}.$$

Hence,

$$V(I) = \{(0,0), (1,1)\}.$$

Since $V(I) \subset \mathbb{R}^2$, we have $V_{\mathbb{R}}(I) = V(I)$. ◇

Recall Definition 3.1.1; in the language of ideals and varieties, the set of steady states of a polynomial dynamical system given by

$$\dot{x}_i = p_i(x),\ i = 1, 2, \ldots, n$$

is precisely

$$V_{\mathbb{R}}(\langle p_1, p_2, \ldots, p_n \rangle).$$

3.4 The division algorithm

In our coming discussion of Gröbner bases, we will need a generalized division algorithm. Let us first recall the situation for univariate polynomials.

Let $k[t]$ be the polynomial ring in one variable over a field k. It is well-known that $k[t]$ is a *principal ideal domain*, i.e. that each ideal $I \subset k[t]$ is generated by a singleton set (see e.g. [4, Corollary 5 in chapter 1]); in other words, that

$$I = \langle p \rangle$$

for some $p \in k[t]$. Let $f \in k[t]$. Then, by the division algorithm for univariate polynomials, there are unique $q, r \in k[t]$ such that

$$f = qp + r$$

with either $r = 0$ or $\deg(r) < \deg(p)$. The polynomial q is called the quotient and r the remainder when dividing f by p. Moreover, the division algorithm for univariate polynomials lets us find such q and r.

We want to generalize the division algorithm in two directions at once: on the one hand, to polynomials in several variables; on the other hand, to more than one divisors. More precisely, we want an algorithm which, given

$$f \in k[x] \text{ and } P = \{p_1, p_2, \ldots, p_m\} \subset k[x],$$

finds unique

$$q_1, q_2, \dots, q_m, r \in k[x]$$

— where r satisfies some appropriate conditions, to be formulated later — such that

$$f = \sum_{i=1}^{m} q_i p_i + r.$$

It turns out that the division algorithm can be generalized in this way — but the q_i and r are unique only if P is a so called *Gröbner basis* of $\langle P \rangle$. When this is the case, the algorithm will produce unique $q_1, q_2, \dots, q_m$ and r, while for general P, on the other hand, it will still produce such $q_1, q_2, \dots, q_m$ and r, but they will not be unique.

We will need the following property of ideals generated by a set of monomials.

Lemma 3.4.1 ([4, Lemma 2 in chapter 2])**.** *Let $M \subset \operatorname{mon}(k[x])$ and let $I = \langle M \rangle$. Then for each $m \in I$, there exists $\tilde{m} \in M$ and $n \in \operatorname{mon}(k[x])$ such that $m = \tilde{m}n$.*

Proof. Take $m \in M$. Then

$$m = \sum_{i=1}^{s} p_i m_i$$

for some $m_i \in M$ and $p_i \in k[x]$. Let

$$\{n_1, n_2, \dots, n_r\} = \operatorname{supp}(\{p_1, p_2, \dots, p_s\}).$$

Then

$$p_i = \sum_{j=1}^{r} c_{ij} n_j$$

for some c_{ij}, so

$$m = \sum_{i=1}^{s} \sum_{j=1}^{r} c_{ij} n_j m_i.$$

Note that $n_j m_i$ is a monomial. Let

$$\{\mu_1, \mu_2, \dots, \mu_k\} = \{n_j m_i \mid 1 \leq j \leq r \text{ and } 1 \leq i \leq r\},$$

i.e. the μ_i are the distinct monomials among the monomials $n_j m_i$. Then

$$m = \sum_{i=1}^{k} d_i \mu_i$$

for some $d_i \in k$. Since $\mu_i \neq \mu_j$ for $i \neq j$, there can be no cancellation in $\sum_{i=1}^{k} d_i \mu_i$. Since m is a monomial, the equality $m = \sum_{i=1}^{k} d_i \mu_i$ can hold if and only if $d_i \neq 0$ for precisely one i, and $d_i = 1$. Thus

$$m = \mu_i = m_k n_j$$

for some k and j. Let

$$\begin{aligned} \tilde{m} &= m_k, \text{ and} \\ n &= n_j \end{aligned}$$

Then $m = \tilde{m}n$. □

We will need a certain property of sequences of monomial ideals, but before we introduce it, let us introduce terminology which will be convenient to use to express that property.

A sequence can have the property that it becomes constant after some index. We will be interested in whether a given sequence of monomials has this property.

Definition 3.4.2. *Let $<$ be a monomial ordering. Let $\mu = (m_i)_{i=1}^{\infty}$ be a sequence of monomials. Let*

$$S(\mu) = \{j \mid \forall i \in N : m_{j+i} = m_j\}$$

and

$$D(\mu) = \begin{cases} \{1, 2, \ldots, \min S(\mu) - 1\}, & \text{if } S(\mu) \neq \emptyset \\ \mathbb{N}, & \text{otherwise} \end{cases}.$$

If $D(\mu) \neq \mathbb{N}$, we say that the sequence is finite. Otherwise, we say that it is infinite.

The value of the sequence changes at least once on $D(\mu)$, but it is constantly $m_{D(\mu)+1}$ on $\mathbb{N} \backslash D(\mu)$.

A sequence can have the property that it is strictly decreasing, in some sense, as the index increases.

Definition 3.4.3. *Let $<$ be a monomial ordering. Let $\mu = (m_i)_{i=1}^{\infty}$ be a sequence of monomials such that $m_i > m_{i+1}$ for all $i \in D(\mu)$. Then we say that μ is a strictly decreasing sequence (with respect to the ordering $<$).*

Before the next lemma, we need the following concept, and a related result.

Definition 3.4.4 ([1, chapter 6])**.** *A chain of ideals $(I_j)_{j=1}^{\infty}$ such that $I_j \subset I_{j+1}$ for all j is called an ascending chain of ideals. If there is a $k \in \mathbb{N}$ such that $I_{k+i} = I_k$ for all $i \in \mathbb{N}$, we say that the chain satisfies the ascending chain condition.*

Definition 3.4.5 ([1, chapter 6])**.** *Let R be a ring such that every ascending chain satisfies the ascending chain condition. Then we say that R is Noetherian.*

We will need the following two well-known results, which we present without proof; the proofs can be found in the referenced book.

Proposition 3.4.6 ([1, Proposition 6.1 and 6.2])**.** *Let R be a ring. The following are equivalent:*

(i) R is Noetherian.

(ii) Every ideal of R is finitely generated.

Proposition 3.4.7 ([1, Corollary 7.6])**.** *Let k be a field. Then $k[x_1, x_2, \ldots, x_n]$ is Noetherian for every $n \geq 1$.*

The following lemma says that a strictly decreasing sequence can not be infinite.

Lemma 3.4.8 ([7, Lemma 5 in chapter 3])**.** *Let $<$ be a monomial ordering. Every strictly decreasing sequence of monomials (with respect to $<$) in $k[x]$ is finite.*

Proof. Let $(m_i)_{i=1}^{\infty}$ be a strictly decreasing sequence of monomials. Let

$$M = \{m_i \mid i \in \mathbb{N}\},$$

and let $I = \langle M \rangle$. Since $k[x]$ is Noetherian, the ideal I is finitely generated; let

$$\langle n_1, n_2, \dots, n_k \rangle = I.$$

Since $I = \langle M \rangle$, Lemma 3.4.1 implies that, for each i, there is an

$$\begin{aligned} &m_i \in M, \text{ and} \\ &\mu_i \in \operatorname{supp}(k[x]) \end{aligned}$$

such that $n_i = m_i \mu_i$. Let

$$\tilde{M} = \{m_i \mid \exists i \exists \mu_i : n_i = m_i \mu_i\} \subset M.$$

Now take $m \in M$. Since

$$\begin{aligned} &M \subset I, \text{ and} \\ &M \subset \operatorname{supp}(k[x]), \end{aligned}$$

there exists $\tilde{m} \in \operatorname{supp}(k[x])$ and an i such that

$$m = \tilde{m} n_i = \tilde{m} m_i \mu_i,$$

by Lemma 3.4.1. This implies that, for all $i \in \{1, 2, \dots, k\}$, we have

$$m_i < m \text{ for all } m \in M.$$

Hence,

$$\min \tilde{M} < m \text{ for all } m \in M$$

and

$$\min \tilde{M} \in M.$$

Since

$$M = \{m_i \mid i \in \mathbb{N}\},$$

there is a K such that

$$m_K = \min \tilde{M}.$$

Since

$$m_K \leq m \text{ for all } m \in M,$$

we must have $m_{K+i} \geq m_K$ for all $i \in \mathbb{N}$. Since the sequence is strictly decreasing, we must have $m_{K+i} = m_K$ for all $i \in \mathbb{N}$. Hence, the sequence is finite. □

For polynomials in one variable, the $m \in \operatorname{supp}(f)$ with maximal degree is usually regarded as the "largest" monomial (in fact, this is the only admissible monomial ordering on the polynomial ring in one variable, since $1 < x$ by one of the conditions on a monomial ordering, which implies that $x < x^2$ by the other condition on a monomial ordering, so, in general, $x^i < x^{i+1}$ for all i). The notion of monomial orderings (see Definition 2.5.7) lets us define the "largest" monomial of a polynomial in several variables.

Definition 3.4.9 ([7, chapter 3.2.2])**.** *Let $<$ be a monomial ordering on $k[x]$. Let $f \in k[x]$ and $\operatorname{supp}(f) = \{m_1, m_2, \ldots, m_r\}$, where the indices are chosen such that $m_1 > m_2 > \cdots > m_r$.*

We say that m_1 is the leading monomial of f. The leading monomial of f will be denoted $\operatorname{lm}(f)$.

If $f = \sum_{i=1}^{r} c_i m_i$, then c_1 is called the leading coefficient of f. The leading coefficient of f will be denoted $\operatorname{lc}(f)$.

Finally, $\operatorname{lt}(f) = \operatorname{lc}(f)\operatorname{lm}(f)$ is called the leading term of f.

Example 3.4.10. Let $p = x_1^2 x_2 x_3^2 + x_1^3 x_2 + x_3^3$.

(a) Let $<$ be any Deglex-ordering. Then $\operatorname{lm}(p) = x_1^2 x_2 x_3^2$, since this monomial has degree 5, while the other monomials in $\operatorname{supp}(p)$ both have lower degrees.

(b) Let $<$ be the Lex-ordering with $x_1 > x_2 > x_3$. Then $\operatorname{lm}(p) = x_1^3 x_2$, since the other monomials in $\operatorname{supp}(p)$ both have only lower powers of x_1.

(c) Let $<$ be one of the Lex-orderings with $x_3 > x_i$ for $i = 1, 2$. Then $\operatorname{lm}(p) = x_3^3$, since the other monomials in $\operatorname{supp}(p)$ both have only lower powers of x_3.

◇

Now we are almost ready to present the division algorithm. Before doing so, we recall the division algorithm for univariate polynomials (with one divisor).

Let $f, g \in k[t]$.

Assume that $\deg(f) < \deg(g)$. Then set $q = 0$ and $r = f$; then $f = qg + r$.

Assume instead that $\deg(f) \geq \deg(g)$. Then we can find $c_1 \in k$ and $m_1 \in \operatorname{mon}(k[t])$ such that

$$\deg(f - c_1 m_1 g) < \deg(f).$$

Let $f_0 = f$. Then define f_{i+1} by the recursive formula

$$f_{i+1} = \begin{cases} f_i - c_i m_i g, & \deg(f_i) \geq \deg(g) \\ f_i, & \text{otherwise} \end{cases}$$

where $c_i \in k$ and $m_i \in \operatorname{mon}(k[t])$ are chosen such that $\deg(f_{i+1}) < \deg(f_i)$, for all i such that $\deg(f_i) \geq \deg(g)$. Since $(\deg(f_i))_{i=1}^{\infty}$ is a strictly decreasing sequence of non-negative numbers and $\deg(g) \geq 0$, there is a k such that $\deg(f_k) < \deg(g)$. Set

$$q = \sum_{i=1}^{k} c_i m_i$$
$$r = f_k.$$

Then $f = qg + r$. In summary, the idea of the division algorithm for univariate polynomials is to, by an appropriate subtraction, successively lower the degree, until the degree of what is left is lower than the degree of the divisor.

The degree of a univariate polynomial is the degree of its leading monomial, so

$$\deg(f) < \deg(g) \text{ if and only if } \deg(\operatorname{lm}(f)) < \deg(\operatorname{lm}(g)).$$

But recall that there is only one ordering of the monomials in mon $(k[t])$, namely

$$t^\alpha < t^\beta \text{ if and only if } \alpha < \beta.$$

Thus

$$\deg(\operatorname{lm}(f)) < \deg(\operatorname{lm}(g)) \text{ if and only if } \operatorname{lm}(f) < \operatorname{lm}(g).$$

This means we could have formulated the previous paragraph in terms of the leading monomials, instead of the degrees, of the polynomials involved. In this language, the idea of the division algorithm for univariate polynomials is to, by an appropriate subtraction, successively lower the leading monomial, until the leading monomial of what is left is smaller than the leading monomial of the divisor.

The previous paragraph suggests what to do for multivariate polynomials and multiple divisiors. Let $f \in k[x]$ and let $G = \{g_1, g_2, \ldots, g_m\} \subset k[x]$. The division algorithm can be described as consisting of major steps, where each major step consists of minor steps. We start with

$$f_{10} = f.$$

If there is a j such that $\operatorname{lm}(g_j) \mid \operatorname{lm}(f_{10})$, we can find $c \in k$ and $m \in \operatorname{mon}(k[x])$ such that

$$\operatorname{lm}(f_{10} - cmg_j) < \operatorname{lm}(f_{10}).$$

Let

$$f_{11} = f_{10} - cmg_j;$$

more generally, given f_{1j}, let

$$f_{1,j+1} = f_{1j} - c_{1j}m_{1j}g_{i_{1j}},$$

where $c_{1j} \in k$ and $m_{1j} \in \operatorname{mon}(k[x])$ are chosen such that $\operatorname{lm}(f_{1,j+1}) < \operatorname{lm}(f_{1j})$, as longs as there exists i such that $\operatorname{lm}(g_i) \mid \operatorname{lm}(f_{1j})$. Defining $f_{1,j+1}$ in terms of f_{1j} constitutes a minor step.

Assume that k is the smallest number for which $\operatorname{lm}(g_i) \nmid \operatorname{lm}(f_{1k})$ for all i. Then we can write

$$f = \sum_{j=1}^{k-1} c_{1j}m_{1j}g_{i_{1j}} + f_{1k}.$$

This ends the first major step. In other words, the first major step ends when the leading monomial of f_{1k} can not be cancelled by subtracting a multiple of one of the g_j. We can not cancel the leading monomial of f_{1k}, but perhaps we can cancel the leading monomial of $f_{1k} - \operatorname{lm}(f_{1k})$. Let

$$f_{20} = f_{1k} - \operatorname{lm}(f_{1k}).$$

Now a new major step begins. It will be shown that number of major steps is finite.

More precisely, the following proposition and its proof gives the division algorithm.

Proposition 3.4.11 ([7, chapter 3.3], [1, Theorem 3 in chapter 2])**.** *Let $f \in k[x]$ and $g_1, g_2, \ldots, g_m \in k[x]$. Fix a monomial ordering $<$. Then there are polynomials $q_1, q_2, \ldots, q_m, r \in k[x]$ (in general not unique) such that*

- $f = \sum_{i=1}^{m} q_i g_i + r$, *and*
- *either* $r = 0$ *or* $\forall i \in \{1, 2, \ldots, m\}$ $\forall m \in \operatorname{supp}(r)$*:* $\operatorname{lm}(g_i) \nmid m$

Proof. Let $f_{10} = f$.

1. Let $I_{1j} = \{i \mid \operatorname{lm}(g_i) \text{ divides } \operatorname{lm}(f_{1j})\}$. For all j such that $I_{1j} \neq \emptyset$, let

   $$f_{1,j+1} = f_{1j} - \operatorname{lc}(f_{1j}) \operatorname{lc}\left(g_{i_{1j}}\right)^{-1} m_{1j} g_{i_{1j}},$$

   where $i_{1j} = \min I_{1j}$ and $\operatorname{lm}(f_{1j}) = \operatorname{lm}\left(g_{i_{1j}}\right) m_{1j}$. Note that $\operatorname{lm}(f_{1,j+1}) < \operatorname{lm}(f_{1j})$, since the leading monomial of $\operatorname{lc}(f_{1j}) \operatorname{lc}\left(g_{i_{1j}}\right)^{-1} m_{1j} g_{i_{1j}}$ is equal to the leading monomial f_{1j}, by construction.

   Assume that there are infinitely many j such that $\operatorname{lm}(g_i) \mid \operatorname{lm}(f_{1j})$ for some i. Then $(\operatorname{lm}(f_{1j})_{j=1}^{\infty}$ is infinite and strictly decreasing, which is a contradiction to Lemma 3.4.8. Hence, there can only be finitely many such j. So

   $$k_1 = \min\{j \mid \forall i : \operatorname{lm}(g_i) \nmid \operatorname{lm}(f_{1j})\} < \infty.$$

   This means

   $$f_{10} = \sum_{j=0}^{k_1-1} \operatorname{lc}(f_{1j}) \operatorname{lc}\left(g_{i_{1j}}\right)^{-1} m_{1j} g_{i_{1j}} + f_{1k_1},$$

   where $\operatorname{lm}(g_i) \nmid \operatorname{lm}(f_{1k_1})$ for all i. Let

   $$r_1 = \operatorname{lc}(f_{1k_1}) \operatorname{lm}(f_{1k_1})$$

   and

   $$f_{20} = f_{1k_1} - r_1.$$

   Note that $\operatorname{lm}(f_{20}) < \operatorname{lm}(f_{1k_1}) < \operatorname{lm}(f_{10})$.

2. Let $I_{2j} = \{i \mid \operatorname{lm}(g_i) \text{ divides } \operatorname{lm}(f_{2j})\}$. For all j such that $I_{2j} \neq \emptyset$, let

   $$f_{2,j+1} = f_{2j} - \operatorname{lc}(f_{2j}) \operatorname{lc}\left(g_{i_{2j}}\right)^{-1} m_{2j} g_{i_{2j}},$$

   where $\operatorname{lm}(f_{2j}) = \operatorname{lm}\left(g_{i_{2j}}\right) m_{2j}$. Note that $\operatorname{lm}(f_{2,j+1}) < \operatorname{lm}(f_{2j})$, since the leading monomial of $\operatorname{lc}(f_{2j}) \operatorname{lc}\left(g_{i_{2j}}\right)^{-1} m_{2j} g_{i_{2j}}$ is equal to the leading monomial f_{2j}, by construction. By the same argument as above,

   $$k_2 = \min\{j \mid \forall i : \operatorname{lm}(g_i) \nmid \operatorname{lm}(f_{2j})\} < \infty.$$

   Then

   $$f_{20} = \sum_{j=0}^{k_2-1} \operatorname{lc}(f_{2j}) \operatorname{lc}\left(g_{i_{2j}}\right)^{-1} m_{2j} g_{i_{2j}} + f_{2k_2}.$$

   Let

   $$r_2 = \operatorname{lc}(f_{2k_2}) \operatorname{lm}(f_{2k_2})$$

   and

   $$f_{30} = f_{2k_2} - r_2.$$

   Note that $\operatorname{lm}(f_{30}) < \operatorname{lm}(f_{2k_2}) < \operatorname{lm}(f_{20})$.

$$\vdots$$

Note that $\operatorname{lm}(f_{i0}) < \operatorname{lm}(f_{i+1,0})$. Hence, $(\operatorname{lm}(f_{i0}))_{i=1}^{\infty}$ is a strictly decreasing sequence, so by Lemma 3.4.8 it is finite. Let

$$s = \min\{j \mid \forall i \in \mathbb{N} : f_{j+i,0} = f_{j0}\}.$$

Then

$$\begin{aligned} f_{20} &= f_{1k_1} - r_1 \\ &= f_{10} - \sum_{j=0}^{k_1-1} \operatorname{lc}(f_{1j}) \operatorname{lc}\left(g_{i_{1j}}\right)^{-1} m_{1j} g_{i_{1j}} - r_1 \end{aligned}$$

so, since $f = f_{10}$,

$$\begin{aligned} f &= f_{20} + \sum_{j=0}^{k_1-1} \operatorname{lc}(f_{1j}) \operatorname{lc}\left(g_{i_{1j}}\right)^{-1} m_{1j} g_{i_{1j}} + r_1 \\ &= f_{30} + \sum_{j=0}^{k_1-1} \operatorname{lc}(f_{1j}) \operatorname{lc}\left(g_{i_{1j}}\right)^{-1} m_{1j} g_{i_{1j}} + \sum_{j=0}^{k_2-1} \operatorname{lc}(f_{2j}) \operatorname{lc}\left(g_{i_{2j}}\right)^{-1} m_{2j} g_{i_{2j}} \\ &\quad + r_1 + r_2 \\ &= \dots \\ &= f_{s0} + \sum_{m=0}^{s-1} \sum_{j=0}^{k_m-1} \operatorname{lc}(f_{mj}) \operatorname{lc}\left(g_{i_{mj}}\right)^{-1} m_{mj} g_{i_{mj}} + \sum_{m=0}^{s-1} r_m. \end{aligned}$$

Assume that $f_{s0} \neq 0$. Then

$$f_{s0} = \sum_{j=0}^{k_s-1} \operatorname{lc}(f_{sj}) \operatorname{lc}\left(g_{i_{sj}}\right)^{-1} m_{sj} g_{i_{sj}} + f_{sk_s}.$$

Let $r_s = \operatorname{lc}(f_{sk_s}) \operatorname{lm}(f_{sk_s})$. Now, $f_{s+1,0}$ is defined by $f_{s+1,0} = f_{sk_s} - r_s$, so

$$\operatorname{lm}(f_{s+1,0}) < \operatorname{lm}(f_{sk_s}) < \operatorname{lm}(f_{s0}),$$

which contradicts that $f_{s+1,0} = f_{s0}$. Hence, $f_{s0} = 0$. This means

$$f = \sum_{m=0}^{s-1} \sum_{j=0}^{k_m-1} \operatorname{lc}(f_{mj}) \operatorname{lc}\left(g_{i_{mj}}\right)^{-1} m_{mj} g_{i_{mj}} + \sum_{m=0}^{s-1} r_m.$$

Recall that $g_{i_{jk}} \in \{g_1, g_2, \dots, g_m\}$ for all j, k. Collect the factors multiplying g_i in a polynomial q_i, and let $r = \sum_{j=1}^{k-1} r_j$. Note that $\forall i \forall j : \operatorname{lm}(g_i) \nmid r_j$, by construction. Thus, $f = \sum_{i=1}^{m} q_i g_i + r$ where r satisfies the conditions of the proposition. □

Definition 3.4.12. *Let $f \in k[x]$ and $G = \{g_1, g_2, \dots, g_m\} \subset k[x]$. Let q_i, $i = 1, 2, \dots, m$, and r be as in Proposition 3.4.11. Then the m-tuple $(q_1, q_2, \dots, q_m)$ is called a quotient, and r is called a remainder — denoted* $\operatorname{rem}(f, G)$ *— of f divided by G.*

As mentioned in the proposition, the $q_1, q_2, \ldots, q_m$ and r are in general not unique (except in the univariate case). This is the reason we speak of "a quotient" and "a remainder" rather than "the quotient" and "the remainder". The source of the non-uniqueness is that in each minor step (see the paragraphs preceding the proposition), there can be more than one j such that $\operatorname{lm}(g_j) \mid \operatorname{lm}(f_{ij})$. Therefore, in each minor step we make a choice, and it turns out that, in general, which quotient and remainder we get depends on the choices we make in the minor steps. In the proof, we chose

$$\min\{i: \ \operatorname{lm}(g_i) \mid \operatorname{lm}(f_{ij})\},$$

but this choice is arbitrary, and, moreover, it depends on the indexation, which itself is arbitrary. However, as was mentioned before, we get unique quotient and remainder if we divide a polynomial f by a set of polynomials G such that G is a Gröbner-basis of $\langle G \rangle$. We will return to this question after the notion of a Gröbner-basis has been introduced.

Let us illustrate the non-uniqueness of the quotient and the remainder with an example.

Example 3.4.13. Let

$$\begin{aligned} f &= x_1^2 + x_2^3, \\ p_1 &= x_1 + x_2^2, \text{ and} \\ p_2 &= x_1^2 + x_2. \end{aligned}$$

Then the method of the proof gives $f = \left(x_1 - x_2^2\right)\left(x_1 + x_2^2\right) + x_2^3 + x_2^4$.

Now reindex the p_i: let

$$\begin{aligned} p_1 &= x_1^2 + x_2, \text{ and} \\ p_2 &= x_1 + x_2^2. \end{aligned}$$

Then the method of the proof gives $f = \left(x_1^2 + x_2\right) + \left(x_2^3 - x_2\right)$.

Thus, we have different quotient and remainder in the two cases. ◇

For a general set $P \subset k[x]$, there is another problem, in addition to the non-uniquness described above: even if $f \in \langle P \rangle$, it can still have a non-zero remainder when divided by P, as the following example shows.

Example 3.4.14. Let

$$f = y^4 - y^2 \in k[x, y]$$

and let $P = \{p_1, p_2\}$ with

$$\begin{aligned} p_1 &= x^2 + y, \text{ and} \\ p_2 &= x^4 - y^4. \end{aligned}$$

Since

$$y^4 - y^2 = (x^2 + y)(x^2 - y) - (x^4 - y^4),$$

we have $f \in \langle P \rangle$.

Let $<$ be the Deglex-ordering with $x > y$. Then

$$\begin{aligned} \operatorname{lm}(p_1) &= x^2, \text{ and} \\ \operatorname{lm}(p_2) &= x^4, \end{aligned}$$

neither of which divides $\operatorname{lm}(f) = y^4$. This implies that y^4 will be a monomial of the remainder. Thus, a remainder of f divided by P is non-zero. $\diamond$

For a general set $P \subset k[x]$, we have only the following implication:

$$\operatorname{rem}(f, P) = 0 \implies f \in \langle P \rangle; \tag{3.1}$$

this holds since if $\operatorname{rem}(f, P) = 0$, we have $f = \sum_{i=1}^{k} p_i q_i$ for some $q_i \in k[x]$, where $\{p_1, p_2, \ldots, p_k\} = P$, so $f \in \langle P \rangle$. However, it turns out that if P is a Gröbner-basis of $\langle P \rangle$, then the reverse implication holds as well. When the notion of a Gröbner basis has been introduced, we will return to this question as well.

The proof of the division algorithm gives us a method for finding a remainder of f divided by the set $\{p_1, p_2, \ldots, p_m\}$. Let us illustrate this with an example.

Example 3.4.15. Let

$$\begin{aligned} f &= 3x_1^3x_2x_3^2 + x_1x_2^2x_3^3 + 2x_1 + x_2^3x_3 - x_3 - 4, \\ p_1 &= x_1^3 + x_3 + 1, \\ p_2 &= x_2^3 - x_3 - 2, \text{ and} \\ p_3 &= x_3 + 2. \end{aligned}$$

Let $<$ be the Lex-ordering with $x_1 > x_2 > x_3$. Let us divide f by $P = \{p_1, p_2, p_3\}$ with respect to this ordering. Let $f_{10} = f$. Since $\operatorname{lm}(f_{10}) = 3x_1^3x_2x_3^2$ is divisble by $\operatorname{lm}(p_1) = x_1^3$, we let

$$\begin{aligned} f_{11} &= f_{10} - 3x_2x_3^2p_1 \\ &= x_1x_2^2x_3^3 + 2x_1 + x_2^3x_3 - 3x_2x_3^3 - 3x_2x_3^2 - x_3 - 4. \end{aligned}$$

Now, $\operatorname{lm}(f_{11}) = x_1x_2^2x_3^3$, which is not divisible by any element in the set $\{\operatorname{lm}(p_1), \operatorname{lm}(p_2), \operatorname{lm}(p_3)\}$. Thus, we set $r_1 = x_1x_2^2x_3^3$, and let

$$\begin{aligned} f_{20} &= f_{11} - r_1 \\ &= 2x_1 + x_2^3x_3 - 3x_2x_3^3 - 3x_2x_3^2 - x_3 - 4. \end{aligned}$$

Again, none of the $\operatorname{lm}(p_i)$ divides $\operatorname{lm}(f_{20}) = 2x_1$. Therefore, we set $r_2 = 2x_1$ and

$$\begin{aligned} f_{30} &= f_{20} - r_2 \\ &= x_2^3x_3 - 3x_2x_3^3 - 3x_2x_3^2 - x_3 - 4. \end{aligned}$$

Now, $\operatorname{lm}(p_1) \nmid \operatorname{lm}(f_{30})$, but $\operatorname{lm}(p_2) \mid \operatorname{lm}(f_{30})$. Therefore, we set

$$\begin{aligned} f_{31} &= f_{30} - x_3p_2 \\ &= -3x_2x_3^3 - 3x_2x_3^2 - x_3 - 4. \end{aligned}$$

Since neither $\operatorname{lm}(p_1)$ nor $\operatorname{lm}(p_2)$ divides $\operatorname{lm}(f_{31})$, but $\operatorname{lm}(p_3) \mid \operatorname{lm}(f_{31})$, we set

$$\begin{aligned} f_{32} &= f_{31} + 3x_2x_3^2p_3 \\ &= 3x_2x_3^2 + x_3^2 - 4. \end{aligned}$$

Again, only $\operatorname{lm}(p_3)$ divides $\operatorname{lm}(f_{32})$, so

$$\begin{aligned} f_{33} &= f_{32} - 3x_2x_3p_3 \\ &= -6x_2x_3 + x_3^2 - 4. \end{aligned}$$

Only $\operatorname{lm}(p_3) \mid \operatorname{lm}(f_{33})$, so

$$\begin{aligned} f_{34} &= f_{33} + 6x_2p_3 \\ &= 12x_2 + x_3^2 - 4. \end{aligned}$$

Now none of the $\operatorname{lm}(p_i)$ divides $\operatorname{lm}(f_{34})$, so we set $r_3 = 12x_2$ and

$$\begin{aligned} f_{40} &= f_{34} - r_3 \\ &= x_3^2 - 4. \end{aligned}$$

Only $\operatorname{lm}(p_3) \mid \operatorname{lm}(f_{40})$, so

$$\begin{aligned} f_{41} &= f_{40} - x_3p_3 \\ &= -2x_3 - 4. \end{aligned}$$

Only $\operatorname{lm}(p_3) \mid \operatorname{lm}(f_{41})$, so

$$\begin{aligned} f_{42} &= f_{41} + 2p_3 \\ &= 0. \end{aligned}$$

This gives

$$\begin{aligned} f &= f_{10} \\ &= f_{11} + 3x_2x_3^2p_1 \\ &= (f_{20} + r_1) + 3x_2x_3^2p_1 \\ &= ((f_{30} + r_2) + r_1) + 3x_2x_3^2p_1 \\ &= (((f_{31} + x_3p_2) + r_2) + r_1) + 3x_2x_3^2p_1 \\ &= \ldots \\ &= 3x_2x_3^2p_1 + x_3p_2 + \left(3x_2x_3 - 3x_2x_3^2 - 6x_2 + x_3 - 2\right)p_3 \\ &\quad\; + \left(x_1x_2^2x_3^3 + 2x_1 + 12x_2\right). \end{aligned}$$

Thus, $f = q_1p_1 + q_2p_2 + q_3p_3 + r$, with

$$\begin{aligned} q_1 &= 3x_2x_3^2 \\ q_2 &= x_3 \\ q_3 &= 3x_2x_3 - 3x_2x_3^2 - 6x_2 + x_3 - 2 \\ r &= x_1x_2^2x_3^3 + 2x_1 + 12x_2. \end{aligned}$$

(Again, the q_i and r are not neccessarily unique, unless P is a Gröbner basis of $\langle P \rangle$ — but it turns out that, in this case, P is a Gröbner basis of $\langle P \rangle$. Hence, the quotient and remainder are unique.) ◇

3.5 Gröbner bases

Let I be an ideal, with a generating set S, i.e. $I = \langle S \rangle$. However, an ideal can have more than one generating set, and depending on what we want to do, it might be more beneficial to work with one generating set rather than another. A Gröbner basis is a generating set with many useful properties.

We will soon define what a Gröbner basis is, but before we can do so, we need to introduce the notion of *the ideal of leading monomials* of an ideal.

Definition 3.5.1 ([7, chapter 3.2.2])**.**

$$\ell(I) = \langle \{\mathrm{lm}(f) \ \mid f \in I\} \rangle$$

is called the ideal of leading monomials of I.

Let P be a set of polynomials and let $I = \langle P \rangle$. Then

$$\langle \{\mathrm{lm}(p) \mid p \in P\} \rangle \subset \ell(I), \tag{3.2}$$

but the inclusion can be proper [7, chapter 3.2.2], as the following example shows.

Example 3.5.2. Let

$$\begin{aligned} p_1 &= x_1 + 2x_2, \text{ and} \\ p_2 &= x_1^2 + x_2, \end{aligned}$$

and let $P = \{p_1, p_2\}$ and $I = \langle P \rangle$. Let $<$ be the Lex-ordering with $x_1 > x_2$. The division algorithm with respect to this ordering gives:

$$\begin{aligned} x_1^2 + x_2 &= (x_1 + 2x_2)(x_1 - 2x_2) + 4x_2^2 + x_2 \\ \Leftrightarrow 4x_2^2 + x_2 &= x_1^2 + x_2 + (2x_2 - x_1)(x_1 + 2x_2) \in I. \end{aligned}$$

Thus,

$$4x_2^2 + x_2 \in I.$$

But

$$4x_2^2 + x_2 \notin \langle x_1 + 2x_2 \rangle,$$

so

$$\langle 4x_2^2 + x_2 \rangle \subsetneq I,$$

which means that both p_1 and p_2 are needed to generate I.

Note that

$$x_2 = (x_1 + 2x_2) - \left(x_1^2 + x_2\right) \in I,$$

so $x_2 \in \mathrm{lm}(I)$. But

$$x_2 \notin \langle x_1 \rangle = \langle x_1,\ x_1^2 \rangle = \langle \mathrm{lm}(p_1), \mathrm{lm}(p_2) \rangle.$$

Thus, $\langle \mathrm{lm}(p_1), \mathrm{lm}(p_2) \rangle \subsetneq \ell(I)$. ◇

Now we can define the notion of a Gröbner basis.

Definition 3.5.3 ([7, Definition 10 in chapter 3])**.** *Let $I \subset k[x]$ be an ideal. Let $\mathcal{G} = \{g_1, g_2, \ldots, g_r\} \subset I$ be a set such that*

$$\ell\left(I\right) = \langle \operatorname{lm}\left(g_1\right), \operatorname{lm}\left(g_2\right), \ldots, \operatorname{lm}\left(g_r\right)\rangle.$$

Then $\mathcal{G}$ is called a Gröbner basis of I.

Remark *One can also define the notion of a reduced Gröbner basis; see e.g. [7, chapter 3.7]. This notion is not neccessary for our purposes.*

Although the definition only requires that the leading monomials of the Gröbner basis generates the initial ideal of I, this turns out to be sufficient for the Gröbner basis to generate I.

Proposition 3.5.4 ([7, Lemma 11 in chapter 3])**.** *Let $\mathcal{G}$ be a Gröbner basis of I. Then $I = \langle \mathcal{G} \rangle$.*

Proof. Let $\mathcal{G} = \{g_1, g_2, \ldots, g_k\}$. Let $f \in I$. Then

$$f = \sum_{i=1}^{k} q_i g_i + r,$$

where q_i and r are as in Proposition 3.4.11. Assume that $r \neq 0$. Since

$$r = f - \sum_{i=1}^{k} q_i g_i \in I,$$

we have $\operatorname{lm}\left(r\right) \in \ell\left(I\right)$. By Lemma 3.4.1, there is an $m \in \operatorname{supp}\left(k[x]\right)$ such that $\operatorname{lm}\left(r\right) = m \operatorname{lm}\left(g_i\right)$ for some $i \in \{1, 2, \ldots, k\}$. But $\operatorname{lm}\left(g_i\right) \nmid \operatorname{lm}\left(r\right)$ for all $i \in \{1, 2, \ldots, k\}$, by the division algorithm, so we have a contradiction. Hence $r = 0$, which implies that $f \in \langle g_1, g_2, \ldots, g_k \rangle$. We conclude that $I \subset \mathcal{G}$.

On the other hand, $\mathcal{G} \subset I$, since every $g_i \in I$, by the definition of a Gröbner-basis. □

As was mentioned earlier, dividing by a set P such that P is a Gröbner-basis of $\langle P \rangle$ gives unique quotient and remainder.

Proposition 3.5.5 (cf. [7, Proposition 14 in chapter 3])**.** *Let $f \in k[x]$ and let $\mathcal{G} = \{g_1, g_2, \ldots, g_s\} \subset k[x]$ be such that $\mathcal{G}$ is a Gröbner-basis of $\langle \mathcal{G} \rangle$. Then the $q_1, q_2, \ldots, q_s, r$ of Proposition 3.4.11 are unique.*

Proof. Assume that

$$f = \sum_{i=1}^{s} q_i g_i + r_1, \text{ and}$$
$$f = \sum_{i=1}^{s} k_i g_i + r_2.$$

Let $I = \langle P \rangle$. Assume that $r_1 \neq r_2$. Then

$$r_1 - r_2 = \sum_{i=1}^{s} (k_i - q_i) g_i \in I,$$

so $\operatorname{lm}(r_1 - r_2) \in \ell(I)$. Since P is a Gröbner basis of I, we have

$$\ell(I) = \langle \operatorname{lm}(g_1), \operatorname{lm}(g_2), \ldots, \operatorname{lm}(g_s) \rangle.$$

By Lemma 3.4.1, this implies that $\operatorname{lm}(g_j) \mid \operatorname{lm}(r_1 - r_2)$ for some j.

Assume $\operatorname{lm}(r_1) > \operatorname{lm}(r_2)$. Then $\operatorname{lm}(r_1 - r_2) = \operatorname{lm}(r_1)$, which implies that $\operatorname{lm}(g_j) \mid \operatorname{lm}(r_i)$. But $\operatorname{lm}(g_j) \nmid m$ for every j and every $m \in \operatorname{supp}(r)$ by Proposition 3.4.11, so this is contradiction.

Assume $\operatorname{lm}(r_2) > \operatorname{lm}(r_1)$; by the same argument as above, this leads to a contradiction.

Hence $\operatorname{lm}(r_1) = \operatorname{lm}(r_2)$. Let

$$\tilde{r}_i = r_i - \operatorname{lt}(r_i).$$

Repeating the same argument for $\operatorname{lm}(\tilde{r}_1)$ and $\operatorname{lm}(\tilde{r}_2)$ shows $\operatorname{lm}(\tilde{r}_1) = \operatorname{lm}(\tilde{r}_2)$. Repeating the same argument as many times as needed — it will need to be repeated only a finite number of times, since the number of monomials of r_1 and r_2 are finite — shows that $r_1 = r_2$. Thus, we have shown that the remainder is unique.

This implies that $\sum_{i=1}^{s} (k_i - q_i) g_i = 0$. Let $p = \sum_{i=1}^{s} (k_i - q_i) g_i$. Let $\operatorname{lm}(g_i) = m_i$ and $\operatorname{lm}(k_i - q_i) = n_i$. Then

$$g_i = c_i m_i + \tilde{p}_i$$

and

$$k_i - q_i = d_i n_i + \tilde{k}_i - \tilde{q}_i$$

for some $c_i, d_i \in k$ and $\tilde{p}_i, \tilde{k}_i - \tilde{q}_i \in k[x]$, where

$$\operatorname{lm}(\tilde{p}_i) < \operatorname{lm}(p_i)$$

and

$$\operatorname{lm}\left(\tilde{k}_i - \tilde{q}_i\right) < \operatorname{lm}(k_i - q_i).$$

Then $p = \sum_{i=1} c_i d_i m_i n_i + h$, where

$$h = \sum_{i=1}^{s} \left(c_i n_i \tilde{p}_i + d_i \left(\tilde{k}_i - \tilde{q}_i\right) m_i + \left(\tilde{k}_i - \tilde{q}_i\right) \tilde{p}_i \right).$$

Since

$$\operatorname{lm}(h) \leq \max\left(\{ n_i \operatorname{lm}(\tilde{p}_i) \} \cup \left\{ \operatorname{lm}\left(\tilde{k}_i - \tilde{q}_i\right) m_i \right\} \cup \left\{ \operatorname{lm}\left(\tilde{k}_i - \tilde{q}_i\right) \operatorname{lm}(\tilde{p}_i) \right\} \right),$$

where $i \in \{1, 2, \ldots, s\}$, and

$$\begin{gathered} n_i \operatorname{lm}(\tilde{p}_i) < n_i m_i, \\ \operatorname{lm}\left(\tilde{k}_i - \tilde{q}_i\right) m_i < n_i m_i, \text{ and} \\ \operatorname{lm}\left(\tilde{k}_i - \tilde{q}_i\right) \operatorname{lm}(\tilde{p}_i) < n_i m_i, \end{gathered}$$

we have $\operatorname{lm}(h) < n_i m_i$. Thus,

$$\operatorname{lm}(p) \in \{n_i m_i \mid i \in \{1, 2, \dots, s\}\},$$

which implies that

$$\operatorname{lm}(p) = \operatorname{lm}(k_i - q_i) \operatorname{lm}(p_i)$$

for some i. Since $p = 0$, we have $\operatorname{lm}(p) = 0$. Thus,

$$\operatorname{lm}(k_i - q_i) \operatorname{lm}(p_i) = 0.$$

Since $\operatorname{lm}(p_i) \neq 0$, this means

$$\operatorname{lm}(k_i - q_i) = 0.$$

But the leading monomial of a polynomial is zero if and only if the polynomial is in fact the zero polynomial. Hence, $k_i = q_i$. We can assume, without loss of generality, that $i = 1$. Then

$$\sum_{i=2}^{s} (k_i - q_i) p_i = 0.$$

Let $\tilde{p} = \sum_{i=2}^{s} (k_i - q_i) p_i$. By repeating the same argument as above $s-1$ times, it follows that $k_i = q_i$ for all i. Thus, we have shown that the quotient is unique. □

When the set we are dividing with is a Gröbner basis of the ideal which the set generates, the converse of (3.1) holds as well.

Corollary 3.5.6 ([7, Proposition 15 in chapter 3])**.** *Let P be a Gröbner basis of $\langle P \rangle$. Then $f \in \langle P \rangle$ if and only if* $\operatorname{rem}(f, P) = 0$*.*

Proof. (3.1) shows that if $\operatorname{rem}(f, P) = 0$, then $f \in \langle P \rangle$. For the conserve, let $P = \{p_1, p_2, \dots, p_2\}$ and let $f \in \langle P \rangle$. Then $f = \sum_{j=1}^{s} q_i p_i$ for some $q_i \in k[x]$. By the uniqueness of the remainder, $\operatorname{rem}(f, P) = 0$. □

3.6 Buchberger's algorithm

For any ideal, a Gröbner basis can be found by using *Buchberger's algorithm.* The algorithm depends on the following concept and result.

Definition 3.6.1 ([7, chapter 3.11])**.** *Let $f, g \in k[x]$. We say that*

$$S(f, g) = \frac{\operatorname{lt}(g) f - \operatorname{lt}(f) g}{\gcd(\operatorname{lm}(f), \operatorname{lm}(g))}$$

is the S-polynomial of f and g.

Convention *If P is a set of polynomials, we will say that the elements in $\{S(p, q) \mid p, q \in P\}$ are the S-polynomials of P.*

Proposition 3.6.2 ([7, Theorem 23 in chapter 3])**.** *The set $G = \{g_1, g_2, \dots, g_s\}$, where* $\operatorname{lc}(g_i) = 1$ *for each i, is a Gröbner basis of $\langle G \rangle$ if and only if*

$$\operatorname{rem}(S(g_i, g_j), G) = 0$$

for all i, j.

Proof. For this proof, the author has followed [4, Theorem 6 in chapter 2], and have also used the proof of [4, Lemma 5 in chapters].

One direction is easy. Let G be a Gröbner basis of $\langle G\rangle$. Since

$$S(g_i, g_j) = \operatorname{lcm}(\operatorname{lm}(g_i),\ \operatorname{lm}(g_j))\left(\frac{1}{\operatorname{lm}(g_i)}g_i - \frac{1}{\operatorname{lm}(g_j)}g_i\right) \in \langle G\rangle,$$

Corollary 3.5.6 implies that $\operatorname{rem}(S(g_i, g_j), G) = 0$.

For the converse, assume that $\operatorname{rem}(S(g_i, g_j), G) = 0$ for all (i, j). We must show that $\ell(\langle G\rangle) = \langle\{\operatorname{lm}(g) \mid g \in G\}\rangle$. Since

$$\langle\{\operatorname{lm}(g) \mid g \in G\}\rangle \subset \ell(\langle G\rangle)$$

in general (see (3.2)), it is enough to show that $\ell(\langle G\rangle) \subseteq \langle\{\operatorname{lm}(g) \mid g \in G\}\rangle$.

Take $m \in \ell(\langle G\rangle)$. Then $m = \operatorname{lm}(f)$ for some $f = \sum_{i=1}^{s} p_i g_i \in \langle G\rangle$. Let $\mu = \max\{\operatorname{lm}(p_i g_i)\}$ and let $J = \{i \mid \operatorname{lm}(p_i g_i) = m\}$. Note that $\operatorname{lm}(f) \leq \mu$. If $\operatorname{lm}(f) = \mu$, we have

$$\operatorname{lm}(f) = \operatorname{lm}(g_j)\operatorname{lm}(p_j) \in \langle\{\operatorname{lm}(g) \mid g \in G\}$$

for some j. If $\operatorname{lm}(f) \neq \mu$, then $\operatorname{lm}(f) < \mu$. Let $J = \{j \mid \operatorname{lm}(p_j g_j) = \mu\}$. Write

$$\begin{aligned} f &= \sum_{j\in J} p_j g_j + \sum_{j\notin J} p_j g_j \\ &= \sum_{j\in J} \operatorname{lt}(p_j)\, g_j + \sum_{j\in J}(p_j - \operatorname{lt}(p_j))\, g_j + \sum_{j\notin J} p_j g_j. \end{aligned}$$

Note that $\operatorname{lm}((p_j - \operatorname{lt}(p_j))\, g_j) < \mu$ for $j \in J$ and $\operatorname{lm}(p_j g_j) < \mu$ for $j \notin J$. Since $\operatorname{lm}(f) < \mu$, we must have that the leading terms in $\sum_{j\in J} \operatorname{lt}(p_j)\, g_j$ cancel, i.e. $\sum_{j\in J} \operatorname{lt}(\operatorname{lt}(p_j)\, g_j) = 0$. Note that

$$\begin{aligned} \operatorname{lt}(\operatorname{lt}(p_j)\, g_j) &= \operatorname{lt}(p_j)\operatorname{lt}(g_j) \\ &= \operatorname{lc}(p_j)\operatorname{lm}(p_j g_j) \\ &= \operatorname{lc}(p_j)\,\mu, \end{aligned}$$

so $\left(\sum_{j\in J} \operatorname{lc}(p_j)\right)\mu = 0$, which implies that $\sum_{j\in J} \operatorname{lc}(p_j) = 0$. By reindexing, we can make $J = \{1, 2, \ldots, l\}$ — therefore, assume, without loss of generality, that $J = \{1, 2, \ldots, l\}$. Then $\sum_{j=1}^{l-1} \operatorname{lc}(p_j) = -\operatorname{lc}(p_l)$. Next, note that $\operatorname{lm}(\operatorname{lt}(p_i)\, g_i) = \operatorname{lm}(p_i)\operatorname{lm}(g_i) = \mu$, so

$$\begin{aligned} &\operatorname{lcm}(\operatorname{lm}(\operatorname{lt}(p_i)\, g_i),\ \operatorname{lm}(\operatorname{lt}(p_j)\, g_j)) \\ =&\operatorname{lcm}(\mu,\ \mu) \\ =&\mu. \end{aligned}$$

This gives

$$\begin{aligned} S(\operatorname{lt}(p_i)\, g_i, \operatorname{lt}(p_j)\, g_j) =&\mu\left(\frac{1}{\operatorname{lt}(\operatorname{lt}(p_i)\, g_i)}\operatorname{lt}(p_i)\, g_i - \frac{1}{\operatorname{lt}(\operatorname{lt}(p_j)\, g_j)}\operatorname{lt}(p_j)\, g_j\right) \\ =&\mu\left(\frac{1}{\operatorname{lt}(p_i)\operatorname{lt}(g_i)}\operatorname{lt}(p_i)\, g_i - \frac{1}{\operatorname{lt}(p_j)\operatorname{lt}(g_j)}\operatorname{lt}(g_j)\, p_j\right) \\ =&\mu\left(\frac{1}{\operatorname{lm}(g_i)}g_i - \frac{1}{\operatorname{lm}(g_j)}g_j\right). \end{aligned}$$

Since $S(g_i, g_j) = L_{ij}\left(\frac{1}{\mathrm{lm}(g_i)}g_i - \frac{1}{\mathrm{lm}(g_j)}g_j\right)$, where $L_{ij} = \mathrm{lcm}\left(\mathrm{lm}\left(g_i\right),\ \mathrm{lm}\left(g_j\right)\right)$, this implies that

$$S(\mathrm{lt}\left(p_i\right)g_i, \mathrm{lt}\left(p_j\right)g_j) = \frac{\mu}{L_{ij}}S(g_i, g_j).$$

However, we also have

$$\begin{aligned} S(\mathrm{lt}\left(p_i\right)g_i, \mathrm{lt}\left(p_j\right)g_j) &= \mu\left(\frac{1}{\mathrm{lc}\left(p_i\right)\mu}\,\mathrm{lt}\left(p_i\right)g_i - \frac{1}{\mathrm{lc}\left(p_j\right)\mu}\,\mathrm{lt}\left(p_j\right)g_j\right) \\ &= \frac{1}{\mathrm{lc}\left(p_i\right)}\,\mathrm{lt}\left(p_i\right)g_i - \frac{1}{\mathrm{lc}\left(p_j\right)}\,\mathrm{lt}\left(p_j\right)g_j \end{aligned}$$

so

$$\begin{aligned} \sum_{i=1}^{l-1}\mathrm{lc}\left(p_i\right)S(\mathrm{lt}\left(p_i\right)g_i, \mathrm{lt}\left(p_l\right)g_l) &= \sum_{i=1}^{l-1}\mathrm{lt}\left(p_i\right)g_i - \left(\sum_{i=1}^{l-1}\mathrm{lc}\left(p_i\right)\right)\frac{1}{\mathrm{lc}\left(p_l\right)}\,\mathrm{lt}\left(p_l\right)g_l \\ &= \sum_{i=1}^{l}\mathrm{lt}\left(p_i\right)g_i \end{aligned}$$

since $\sum_{i=1}^{l-1}\mathrm{lc}\left(p_i\right) = -\,\mathrm{lc}\left(p_l\right)$. Hence

$$\sum_{i=1}^{l}\mathrm{lt}\left(p_i\right)g_i = \sum_{i=1}^{l-1}\mathrm{lc}\left(p_i\right)\frac{\mu}{L_{il}}S(g_i, g_l).$$

By assumption, $\mathrm{rem}(S(g_i, g_l), G) = 0$ for all i. This means

$$S(g_i, g_l) = \sum_{m=1}^{s} q_m^{(i,l)}g_m,$$

for some $q_m^{(i,l)} \in k[x]$. This gives

$$\sum_{i=1}^{l}\mathrm{lt}\left(p_i\right)g_i = \sum_{m=1}^{s}\tilde{p}_m g_m$$

where

$$\tilde{p}_m = \sum_{i=1}^{l-1}\mathrm{lc}\left(p_i\right)\frac{\mu}{\mathrm{lcm}\left(\mathrm{lm}\left(g_i\right),\ \mathrm{lm}\left(g_j\right)\right)}q_m^{(i,l)}.$$

Thus,

$$\begin{aligned} &\mathrm{lm}\left(\tilde{p}_m g_m\right) \\ &\leq \max\left\{\mathrm{lm}\left(\mathrm{lc}\left(p_i\right)\right)\frac{\mu}{\mathrm{lcm}\left(\mathrm{lm}\left(g_i\right),\ \mathrm{lm}\left(g_j\right)\right)}q_m^{(i,l)}g_m \mid i \in \{1, 2, \ldots, l\}\right\}. \end{aligned}$$

Note that

$$\mathrm{lm}\left(\mathrm{lc}\left(p_i\right)\frac{\mu}{L_{il}}q_m^{(i,l)}g_m\right) = \mathrm{lm}\left(\frac{\mu}{L_{il}}\right)\mathrm{lm}\left(q_m^{(i,l)}g_m\right).$$

By the division algorithm, we have $\mathrm{lm}\left(q_m^{(i,l)}g_m\right) \leq \mathrm{lm}\left(S(g_i, g_j)\right)$. Also,

$$\mathrm{lm}\left(\frac{\mu}{L_{il}}\right) = \frac{\mu}{L_{il}},$$

since $\frac{\mu}{L_{il}}$ is a monomial. This implies that

$$\operatorname{lm}\left(\operatorname{lc}(p_i)\,\frac{\mu}{L_{il}}q_m^{(i,l)}g_m\right) \leq \frac{\mu}{L_{il}}S(g_i, g_j).$$

Now, write $g_i = \operatorname{lt}(g_i) + \tilde{g}_i$ and $g_i = \operatorname{lt}(g_j) + \tilde{g}_j$. Then

$$S(g_i, g_j) = \frac{L_{il}}{\operatorname{lm}(g_i)}\tilde{g}_i - \frac{L_{il}}{\operatorname{lm}(g_j)}\tilde{g}_j,$$

which implies that

$$\operatorname{lm}(S(g_i, g_l)) \leq \max\left\{\operatorname{lm}\left(\frac{L_{il}}{\operatorname{lm}(g_i)}\tilde{g}_i\right), \operatorname{lm}\left(\frac{L_{il}}{\operatorname{lm}(g_l)}\tilde{g}_l\right)\right\}.$$

Note that

$$\operatorname{lm}\left(\frac{L_{il}}{\operatorname{lm}(g_i)}\operatorname{lm}(\tilde{g}_i)\right) < \frac{L_{il}}{\operatorname{lm}(g_i)}\operatorname{lm}(g_i) = L_{il},$$

where we have used that $\operatorname{lm}(\tilde{g}_i) < \operatorname{lm}(g_i)$. In the same way,

$$\operatorname{lm}\left(\frac{L_{il}}{\operatorname{lm}(g_l)}\tilde{g}_l\right) < L_{il}.$$

Thus, $\operatorname{lm}(S(g_i, g_l)) < L_{il}$, so

$$\operatorname{lm}\left(\operatorname{lc}(p_i)\,\frac{\mu}{L_{il}}q_m^{(i,l)}g_m\right) < \mu,$$

which implies that $\operatorname{lm}(\tilde{p}_m g_m) < \mu$. Thus,

$$\sum_{i=1}^{s} p_i g_i = \sum_{j\in J}\tilde{p}_j g_j + \sum_{j\in J}(p_j - \operatorname{lt}(p_j))\,g_j + \sum_{j\notin J} p_j g_j,$$

where each summand has leading monomial strictly smaller than μ.

Let

$$\sum_{i=1}^{s} p_i g_i = \sum_{j\in J}\tilde{p}_j g_j + \sum_{j\in J}(p_j - \operatorname{lt}(p_j))\,g_j + \sum_{j\notin J} p_j g_j = \sum_{i=1}^{s} P_i g_i,$$

and let $\tilde{\mu} = \max\{\operatorname{lm}(P_i g_i) \mid i \in \{1, 2, \ldots, s\}\}$. If $\operatorname{lm}(f) = \tilde{\mu}$, we have $\operatorname{lm}(f) = \operatorname{lm}(P_i)\operatorname{lm}(g_i) \in \langle\{\operatorname{lm}(g) \mid g \in G\}\rangle$. Otherwise, $\operatorname{lm}(f) < \tilde{\mu}$; if so, repeat the procedure above.

Let $\mu_0 = \mu$, $\mu_1 = \tilde{\mu}$, and let μ_j be the maximum leading monomial after the procedure has been repeated j times. Then $(\mu_j)_{j=1}^{\infty}$ is a strictly decreasing sequence of monomials, so by Lemma 3.4.8, it is is finite, i.e. there is a k such that $\mu_{k+i} = \mu_k$ for all $i \geq 0$. Assume that $\mu_k > \operatorname{lm}(f)$: then we could repeat the procedure above to get $\mu_{k+1} < \mu_k$, which is a contradiction. Hence $\mu_k \leq \operatorname{lm}(f)$. Since $\operatorname{lm}(f) \leq \mu_j$ for all j, this implies that $\operatorname{lm}(f) = \mu_k$. Thus,

$$\operatorname{lm}(f) = \operatorname{lm}(\hat{p}_i)\operatorname{lm}(g_i) \in \langle\{\operatorname{lm}(g) \mid g \in G\}\rangle$$

for some $\operatorname{lm}(\hat{p}_i)$.

Thus, we have shown that

$$m = \operatorname{lm}(f) \in \langle\{\operatorname{lm}(g) \mid g \in G\}\rangle,$$

so $\ell(\langle G\rangle) \subset \langle\{\operatorname{lm}(g) \mid g \in G\}\rangle$. Hence, G is a Gröbner basis of $\langle G\rangle$. □

Consider $\langle P\rangle \subset k[x]$. The idea of Buchberger's algorithm is as follows. The set P is our candidate for a Gröbner basis. By Proposition 3.6.2, it is a Gröbner basis if and only if $\operatorname{rem}(S(p_i, p_j), P) = 0$ for all $p_i, p_j \in P$. Extend the set P with all non-zero remainders — let us call the extended set P' — and then start over with P' as the new candidate for a Gröbner basis. Note that $\langle P'\rangle = \langle P\rangle$, since each remainder is given by

$$r = S(p_i, p_j) - \sum_{i=1}^{s} q_i p_i \in \langle P\rangle.$$

Repeat until all remainders of all S-polynomials of the candidate set are zero; then the candidate set is a Gröbner basis of $\langle P\rangle$.

More formally, we have the following proposition.

Proposition 3.6.3 (Buchberger's algorithm; based on [7, chapter 3.13] and [4, chapter 3.7]). *Let $P = \{p_1, p_2, \dots, p_k\}$ and let $I = \langle P\rangle$. Let $G_0 = P$. Let*

$$S_j = \{\operatorname{rem}(S(p,q), G_{j-1}) \mid p, q \in G_{j-1} \text{ and } \operatorname{rem}(S(p,q), G_{j-1}) \neq 0\}, \text{ and}$$
$$G_j = G_{j-1} \cup S_j$$

Let $k = \min\{j \mid S_j = \emptyset\}$. Then $k < \infty$, and G_{k-1} is a Gröbner-basis of I.

Proof. Assume that $S_j \neq \emptyset$ for all $j \in \mathbb{N}$. Then

$$G_0 \subsetneq G_1 \subsetneq G_2 \subsetneq \dots.$$

Take $r \in G_{i+1} \backslash G_i$ for some i. Since

$$G_{i+1} = G_i \cup S_{i+1},$$

this means $r = \operatorname{rem}(S(p,q), G_i)$ for some $p, q \in G_i$. By the division algorithm, this means $\operatorname{lm}(g) \nmid \operatorname{lm}(r)$ for all $g \in G_i$.

Assume that $\operatorname{lm}(r) \in \ell(G_i)$. Then Lemma 3.4.1 implies $\operatorname{lm}(r) = m \operatorname{lm}(g)$ for some $g \in G_i$, so $\operatorname{lm}(g) \mid \operatorname{lm}(r)$. This is a contradiction to $\operatorname{lm}(g) \nmid \operatorname{lm}(r)$ for all $g \in G_i$.

Hence, $\operatorname{lm}(r) \notin \ell(G_i)$. But $\operatorname{lm}(r) \in \ell(G_{i+1})$ since $r \in G_{i+1}$. Hence

$$\ell(G_i) \subsetneq \ell(G_{i+1}).$$

This implies that $(\ell(G_i))_{i=1}^{\infty}$ is an ascending chain of ideals, so by Proposition 3.4.7 it satisfies the ascending chain condition. Hence, there is a k such that $G_k = G_{k-1}$, which implies that $S_k = \emptyset$.

If $S_k = \emptyset$, then $\operatorname{rem}(S(p,q), G_{k-1}) = 0$ for all $p, q \in G_{k-1}$. By Proposition 3.6.2, this is equivalent to G_{k-1} being a Gröbner basis of I. □

Example 3.6.4. Let

$$p_1 = x^2 + 1, \text{ and}$$
$$p_2 = y^2 + 1.$$

Let $P = \{p_1, p_2\} \subset k[x, y]$. Let $<$ be the Lex-ordering with $x > y$. Since

$$\begin{aligned} S(p_1, p_2) &= -x^2 + y^2 \\ &= -p_1 + p_2 \end{aligned}$$

we have $\operatorname{rem}(S(p_1, p_2), P) = 0$. Thus, P is a Gröbner-basis of $\langle P\rangle$. ⋄

Example 3.6.5. Let

$$\begin{aligned} p_1 &= x^2 + 1, \text{ and} \\ p_2 &= x + y^2 \end{aligned}$$

Let $P = \{p_1, p_2\} \subset k[x, y]$. Let $<$ be the Lex-ordering with $x > y$. The S-polynomial of p_1 and p_2 is

$$S(p_1, p_2) = -xy^2 + 1.$$

The division algorithm gives

$$S(p_1, p_2) = -y^2 p_2 + (y^4 + 1),$$

where neither y^4 nor 1 is divisible by any element in

$$\{\operatorname{lm}(p_1), \operatorname{lm}(p_2)\} = \left\{x^2,\ x\right\}.$$

Thus, $\operatorname{rem}(S(p_1, p_2), P) = y^4 + 1$. Hence, P is not a Gröbner-basis of P.

Let $p_3 = y^4 + 1$ and let $P' = P \cup \{p_3\}$. Then $S(p_1, p_2) = -y^2 p_2 + p_3$. The other S-polynomials are

$$\begin{aligned} S(p_1, p_3) &= -x^2 + y^4, \text{ and} \\ S(p_2, p_3) &= -x + y^6 \end{aligned}$$

The division algorithm gives $S(p_1, p_3) = p_3 - p_1$ and $S(p_2, p_3) = y^2 p_3 - p_2$. Thus, $\operatorname{rem}(S(p_i, p_3), P') = 0$ for $i = 1, 2$. Hence, P' is a Gröbner-basis for $\langle P' \rangle = \langle P \rangle$. ◇

3.7 Strongly triangular form

Let I be an ideal such that its variety consists of finitely many elements. Then any Gröbner basis of I will have a useful structure: the Gröbner basis has a so called *strongly triangular form*, which allows us to find $V(I)$ — from which the set of steady states, i.e. $V_{\mathbb{R}}(I)$, is immediately found by taking the intersection of $V(I)$ with $\mathbb{R}$ — by finding the roots of a sequence of univariate polynomials.

First, let us introduce the notion of a *triangular form* on a set of polynomials; first informally, then followed by a precise definition. Let $P \subset k[x_1, x_2, \ldots, x_n]$ be a set. An arbitrary $p \in P$ might not depend on all variables; in other words, it might be that not all variables $x_1, x_2, \ldots, x_n$ appear in p.

Example 3.7.1. Let

$$P = \left\{x_1 x_3,\ x_2 x_3^2,\ x_1^3 x_2 x_3^2\right\} \subset k[x_1, x_2, x_3].$$

In $x_1 x_3$ the variables x_1 and x_3, but not x_2, appear. In $x_2 x_3^2$ the variables x_2 and x_3, but not x_1, appear. In $x_1^3 x_2 x_3^2$ all variables in $k[x_1, x_2, x_3]$ appear. ◇

With this in mind, let us order the polynomials of P in the following order:

1. The polynomials in which x_1 appear.

2. The polynomials in which x_2, but not x_1, appear.

3. The polynomials in which x_3, but neither x_1 nor x_2, appear.

$\vdots$

i. The polynomials in which x_i, but none of $x_1, x_2, \ldots, x_{i-1}$, appear.

$\vdots$

n. The polynomials in which x_n, but none of $x_1, x_2, \ldots, x_{n-1}$, appear.

n+1. The constant polynomials.

Within each group, the order does not matter.

An ordering of polynomials as above (not to be confused with a monomial ordering) is called a triangular form on P. Let us make a precise definition.

Definition 3.7.2 ([11, Definition 4.2.1])**.** *Let $P \subset k[x]$. Let $\sigma(j) = i_j,\ j = 1, 2, \ldots, n$. Let*

$$P_m = \begin{cases} P \cap \left(k[x_{i_{m+1}}, x_{i_{m+2}}, \ldots, x_{i_n}] \backslash k[x_{i_{m+2}}, \ldots, x_{i_n}]\right), & m = 0, 1, \ldots, n-2 \\ P \cap \left(k[x_{i_n}] \backslash k\right), & m = n-1 \\ P \cap k, & m = n \end{cases}.$$

Let $<$ be an order on P such that for all m the following holds: if $p_m \in P_m$ and $p_{m+1} \in P_{m+1}$, then $p_m < p_{m+1}$. Then $<$ is called a triangular form on P.

Remark *We use $<$ for both monomial orderings and for triangular forms. We will be precise about in which sense $<$ is used.*

A triangular form is just a special partition of the set. Thus, there is a triangular form on every set of polynomials. However, not every triangular form is in *strongly triangular form*, which we now define.

Definition 3.7.3 (cf. [11, Definition 4.2.2])**.** *Let $P \subset k[x]$ be a set and let $<$ be a triangular form on P. Assume that $P \cap k = \emptyset$ and such that, for some σ, it holds that for every i, there is a $p_i \in P_i$, where P_i is defined as in Definition 3.7.2, with $x_i^{\alpha_i} \in \operatorname{supp}(p_i)$ for some $\alpha_i \in \mathbb{R}$. Then we say that $<$ is in strongly triangular form.*

Whether a triangular form on a set P is in strongly triangular form or not depends only on P itself. We make the following convention.

Convention *If there is a triangular form $<$ on P, such that $<$ is in strongly triangular form, we will say that P is strongly triangular.*

Not every set of polynomials admits a strongly triangular form. A trivial counterexample is any set which includes a constant polynomial. An example of a set which does not include any constant polynomials, but still does not admit a strongly triangular form, is $\{x_1x_2,\ x_3\} \subset k[x_1, x_2, x_3]$; the set $\{x_1x_2,\ x_3\} \cap (k[x_2, x_3] \backslash k[x_3])$ is empty.

Note that, if a set P is strongly triangular, then, in particular, $p_i \in P_i$ for every i, where P_i is as in Definition 3.7.2.

3.8 An algorithm for finding the steady states

It turns out that among the sets P such that P is a Gröbner basis of $\langle P\rangle$, the sets which are strongly triangular are precisely the sets with $|V(\langle P\rangle)| < \infty$. For the proof of this, we need some results from [7].

Proposition 3.8.1 ([7, chapter 3.6]). *Let $<$ be any monomial ordering. Then $\mathcal{B} = \operatorname{mon}(k[x]) \setminus \ell(I)$ is a basis for the k-vector space $k[x]/I$.*

Proof. Let $I = \{p_1, p_2, \ldots, p_k\}$. Let $f \in k[x]$. Then $f = \sum_{i=1}^{k} q_i p_i + r$, with $r = \sum_{\alpha \in A} c_\alpha m_\alpha$, where $A = \{\alpha \mid m_\alpha \in \mathcal{B}\}$. This shows immediately that $\operatorname{span}(\mathcal{B}) = k[x]/I$.

Assume that there exists $B \subset A$ and $\left\{c_\alpha^B \mid \alpha \in B\right\} \neq \{0\}$ such that $\sum_{\alpha \in B} c_\alpha^B m_\alpha \in I$. Then

$$m_\gamma = \operatorname{lm}\left(\sum_{\alpha \in B} c_\alpha^B m_\alpha\right) \in \ell(I)$$

for some $\gamma \in B$, which contradicts that $m_\alpha \notin \ell(I)$ for every $\alpha \in B$. Thus, $\mathcal{B}$ is linearly independent in $k[x]/I$. □

Lemma 3.8.2 ([7, Corollary 5 in chapter 6.2]). *If $|V(I)| < \infty$, then $\mathbb{C}[x]/I$ is finite-dimensional as a vector space over $\mathbb{C}$.*

Remark *The converse holds as well. [7, Theorem 7]*

Lemma 3.8.3 (one part of [7, Proposition 6 in chapter 6.2]). *Let $I \subset \operatorname{mon}(k[x])$. If $\mathbb{C}[x]/I$ is finite-dimensional as a $\mathbb{C}$-vector space then, for every $i \in \{1, 2, \ldots, n\}$, there is an $m_i \in \mathbb{N}$ such that $x_i^{m_i} \in I$.*

Remark *The converse holds as well; this is the second part of the cited proposition.*

Proposition 3.8.4. *Let $I \subset \mathbb{C}[x]$ be an ideal. Let $<$ be any Lex-ordering. Then $|V(I)| < \infty$ if and only if every Gröbner basis of I, with respect to $<$, is strongly triangular.*

Proof. Assume that every Gröbner basis of I, with respect to $<$, is strongly triangular and let $\mathcal{G}$ be a Gröbner basis of I with respect to this ordering. The proof of this implication is based on the algorithm described in [7, chapter 6.3] (although, there, the argument is not phrased in terms of strongly triangular forms). We can, without loss of generality, assume that $x_1 > x_2 > \ldots x_n$. Since $\mathcal{G}$ is strongly triangular, there are polynomials $\{g_1, g_2, \ldots, g_n\} \subset \mathcal{G}$ such that

$$\begin{aligned}
g_n &\in \mathbb{C}[x_n],\\
g_{n-1} &\in \mathbb{C}[x_{n-1}, x_n] \backslash \mathbb{C}[x_n],\\
g_{n-2} &\in \mathbb{C}[x_{n-2}, x_{n-1}, x_n] \backslash \mathbb{C}[x_{n-1}, x_n],\\
&\vdots\\
g_2 &\in \mathbb{C}[x_2, x_3, \ldots, x_{n-1}, x_n] \backslash \mathbb{C}[x_3, x_4, \ldots, x_{n-1}, x_n], \text{ and}\\
g_1 &\in \mathbb{C}[x_1, x_2, x_3, \ldots, x_{n-1}, x_n] \backslash \mathbb{C}[x_2, x_3, x_4, \ldots, x_{n-1}, x_n].
\end{aligned}$$

Since g_n is a univariate polynomial, it has finitely many solutions. Let

$$C_n = \{\alpha_n \in \mathbb{C} \mid g_n(\alpha_n) = 0\}\,;$$

then $|C_n| < \infty$. Define recursively

$$C_i = \Big\{(\alpha_i, \alpha_{i+1}, \ldots, \alpha_n) \in \mathbb{C}^{n-(i-1)} \mid (\alpha_{i+1}, \alpha_{i+2}, \ldots, \alpha_n) \in C_{i+1} \text{ and } g_i(\alpha_i, \alpha_{i+1}, \ldots, \alpha_n) = 0\}$$

for $i = n-1, n-2, \ldots, 2, 1$. At the j:th step, we find the roots of

$$g_{n-(j-1)}(x_{n-(j-1)}, \alpha_{n-(j-2)}, \alpha_{n-(j-3)}, \ldots, \alpha_n).$$

This is a univariate polynomial, so it has finitely many solutions. Since C_n is finite, this implies that C_{n-1} is finite, which in turn implies that C_{n-2} is finite, and so on; hence, each C_i is finite. In particular, C_1 is finite. It is clear, from the way it was constructed, that $C_1 = V(I)$. Thus, $|V(I)| < \infty$.

Assume that $|V(I)| < \infty$ and let $\mathcal{G}$ be a Gröbner basis of I with respect to $<$. For the proof of this implication, the author has followed the reasoning in [7, p. 83–84]. Lemma 3.8.2 implies that $\mathbb{C}[x]/I$ is finite-dimensional as a $\mathbb{C}$-vector space. Proposition 3.8.1 implies that the $\mathbb{C}$-vector spaces $\mathbb{C}[x]/I$ and $\mathbb{C}[x]/\,\ell\,(I)$ have the same dimension. Since $\ell\,(I)$ is a monomial ideal, Lemma 3.8.3 implies that for every i there is an $m_i \in \mathbb{N}$ such that $x_i^{m_i} \in \ell\,(I)$. Again, we can, without loss of generality since we can always re-index the variables, assume that

$$x_1 > x_2 > \cdots > x_n.$$

There is an $m_n \in N$ such that $x_n^{m_n} \in \ell\,(I)$. Since $\mathcal{G}$ is a Gröbner basis of I, we have $\ell\,(I) = \langle\{\mathrm{lm}\,(g) \mid g \in \mathcal{G}\}\rangle$. It follows from Lemma 3.4.1 that

$$x_n^{m_n} = \mathrm{lm}\,(g_n)\,M_n$$

for some monomial M_n and some $g_n \in \mathcal{G}$. This implies that $\mathrm{lm}\,(g_n) = x_n^{\tilde{m}_n}$ and $M_n = x_n^{k_n}$ for some $\tilde{m}_n, k_n \in \mathbb{N}$ (such that $m_n = \tilde{m}_n + k_n$). Since x_n is the smallest variable, we must have that $g_n \in \mathbb{C}[x_n]$.

There is also an $m_{n-1} \in \mathbb{N}$ such that $x_{n-1}^{m_{n-1}} \in \ell\,(I)$. By the same reasoning, there is a $g_{n-1} \in \mathcal{G}$ such that $\mathrm{lm}\,(g_{n-1}) = x_{n-1}^{\tilde{m}_{n-1}}$ for some $\tilde{m}_{n-1} \in \mathbb{N}$. Since x_{n-1} is the next to smallest variable, this can hold only if $g_{n-1} \in \mathbb{C}[x_{n-1}, x_n]$. Also note that $g_{n-1} \notin \mathbb{C}[x_n]$, since $x_{n-1}^{\tilde{m}_{n-1}}$ is a term of g_{n-1}. Repeating the same argument $(n-2)$ more times, we find

$$g_i \in \mathcal{G} \cap (\mathbb{C}[x_i, x_{i+1}, \ldots, x_n] \backslash \mathbb{C}[x_{i+1}, x_{i+2}, \ldots, x_n])$$

for $i = 3, 4, \ldots, n$ as well. Thus, we have shown that $\mathcal{G}$ is strongly triangular. □

Let $\dot{x}_i = p_i(x)$, $i = 1, 2, \ldots, n$, be a polynomial dynamical system. Let $P = \{p_1, p_2, \ldots, p_n\}$. If $|V(\langle P\rangle)| < \infty$, then the proof of Proposition 3.8.4 gives an algorithm for finding the steady states of the system: since $V(\langle P\rangle) = C_1$ and the set of steady states of the system is equal to $V_{\mathbb{R}}(\langle P\rangle) = V(\langle P\rangle) \cap \mathbb{R}^n$, the set of steady states is equal to $C_1 \cap \mathbb{R}^n$ (where C_1 is defined as in the proof). For convenience, let us record the algorithm in a corollary to Proposition 3.8.4. For convenience of notation, let "the proposition" be short for "Proposition 3.8.4" in the formulation of the corollary.

Corollary 3.8.5. *Let $\dot{x}_i = p_i(x),\ i = 1, 2, \ldots, n$ be a polynomial dynamical system such that $|V(\langle p_1, p_2, \ldots, p_n\rangle)| < \infty$.*

1. *Compute a Gröbner basis $\mathcal{G}$ of $\langle p_1, p_2, \ldots, p_n\rangle$. By the proposition, $\mathcal{G}$ has strongly triangular form.*
2. *Let $g_n \in \mathcal{G} \cap (\mathbb{C}[x_n])$; such a g_n exists, by the proposition. Solve $g_n(x) = 0$. Let C_n be the set of roots of g_n.*
3. *Let $g_{n-1} \in \mathcal{G} \cap (\mathbb{C}[x_{n-1}, x_n] \backslash \mathbb{C}[x_n])$; such a g_{n-1} exists, by the proposition. For each $\alpha_n \in \mathbb{C}_n$, solve $g(x_{n-1}, \alpha_n) = 0$. Let C_{n-1} be the set of roots (α_{n-1}, α_n) of g_{n-1} such that $\alpha_n \in C_n$.*
4. *Let $g_{n-2} \in \mathcal{G} \cap (\mathbb{C}[x_{n-2}, x_{n-1}, x_n] \backslash \mathbb{C}[x_{n-1}, x_n])$; such a g_{n-2} exists, by the proposition. For each $(\alpha_{n-1}, \alpha_n) \in C_{n-1}$, solve $g_{n-2}(x_{n-2}, \alpha_{n-1}, \alpha_n) = 0$, Let C_{n-2} be the set of roots $(\alpha_{n-2}, \alpha_{n-1}, \alpha_n)$ of g_{n-2} such that $(\alpha_{n-1}, \alpha_n) \in C_{n-1}$.*

$\vdots$

n. Let $g_1 \in \mathcal{G} \cap (\mathbb{C}[x_1, x_2, \ldots, x_n] \backslash \mathbb{C}[x_2, x_3, \ldots, x_n])$; such a g_1 exists, by the proposition. For each $(\alpha_2, \alpha_3, \ldots, \alpha_n) \in C_2$, solve $g_1(x_1, \alpha_2, \alpha_3, \ldots, \alpha_n) = 0$. Let C_1 be the set of roots $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ of g_1 such that $(\alpha_2, \alpha_3, \ldots, \alpha_n) \in C_2$.

Then the set of steady states is equal to $C_1 \cap \mathbb{R}^n$.

Example 3.8.6. Consider the system $\dot{x}_i = p_i(x),\ i = 1, 2, 3$, where

$$\begin{aligned} p_1 &= x_1^2 x_2^3 x_3 - x_1(x_2 x_3^2 + x_2^2 x_3) + x_2 x_3^3, \\ p_2 &= x_2^2 x_3^4 + 2x_2 x_3^2 + x_3^6 - 4, \text{ and} \\ p_3 &= x_2 x_3^2 - 1. \end{aligned}$$

The set $\mathcal{G} = \{g_1, g_2, g_3\}$, with

$$\begin{aligned} g_1 &= x_1^2 - x_1 x_3^5 - x_1 x_3^2 + 1, \\ g_2 &= x_2 - x_3^4, \text{ and} \\ g_3 &= x_3^6 - 1, \end{aligned}$$

is a Gröbner basis, with respect to the Lex-ordering with $x_1 > x_2 > x_3$, for the ideal $\langle p_1, p_2, p_3\rangle$. Note that

$$\begin{aligned} g_3 &\in \mathbb{R}[x_3], \\ g_2 &\in \mathbb{R}[x_2, x_3] \backslash \mathbb{R}[x_3], \text{ and} \\ g_1 &\in \mathbb{R}[x_1, x_2, x_3] \backslash \mathbb{R}[x_2, x_3]. \end{aligned}$$

Thus, $\mathcal{G}$ is strongly triangular.

The real solutions of $x_3^6 - 1 = 0$ are $x_3 = \pm 1$. In either case, we get $g_2 = x_2 - 1$, which has the solution $x_2 = 1$. If $x_3 = 1$, then $g_1 = x_1^2 - 2x_1 + 1 = (x_1 - 1)^2$, so $x_1 = 1$. If $x_3 = -1$, then $g_1 = x_1^2 + 1$, which has no real solution. Thus, the only steady state of the system is $(1, 1, 1)$. ◇

4 Reduction of the number of parameters of a system

Let $\dot{X} = P(X) = (p_1(X), \ldots, p_n(X))$ be a polynomial dynamical system (we use X now so that we can use x later). Let $P = \{p_1, p_2, \ldots, p_n\}$. Then $p_i = \sum_{m_\alpha \in \operatorname{supp}(P)} c_{i,\alpha} m_\alpha$ for some $c_{i,\alpha} \in \mathbb{R}$. Sometimes some or all of the $c_{i,\alpha}$ are considered as, or depends on, parameters; e.g. the k_i and k_{-i} in (7.3). We now present a way to possibly reduce the number of parameters in the system. Let $\tau = \sigma t$, where $\sigma \in \mathbb{R}$, and let $x_i(\tau) = \xi_i X_i(\tau/\sigma)$, where $\xi_i \in \mathbb{R}$, for $i = 1, 2, \ldots, n$. Then

$$\frac{dx_i}{d\tau} = \sum_{m_\alpha \in \operatorname{supp}(P)} \frac{\xi_i c_{i,\alpha}}{\sigma \xi^{\operatorname{mdeg}(m_\alpha)}} m_\alpha, \tag{4.1}$$

where $\xi^{\operatorname{mdeg}(m_\alpha)} = (\xi_1^{\alpha_1}, \xi_2^{\alpha_2}, \ldots, \xi_n^{\alpha_n})$ with $\operatorname{mdeg}(m_\alpha) = (\alpha_1, \alpha_2, \ldots, \alpha_n)$. Equation (4.1) suggests a natural way to get rid of a parameter $c_{i,\alpha}$: choose $\xi = (\xi_1, \xi_2, \ldots, \xi_n)$ so that

$$\frac{\xi_i c_{i,\alpha}}{\sigma \xi^{\operatorname{mdeg}(m_\alpha)}} = 1.$$

If $c_{i,\alpha} = 0$, this equation has no solution. Let $A_i = \{\alpha \mid m_\alpha \in \operatorname{supp}(P),\ c_{i,\alpha} \neq 0\}$. We are led to the system of equations

$$\sigma \xi^{\operatorname{mdeg}(m_\alpha)} - c_{i,\alpha}\xi_i = 0,\ \alpha \in A_i,\ i = 1, 2, \ldots, n. \tag{4.2}$$

Note that $\xi^{\operatorname{mdeg}(m_\alpha)} - c_{i,\alpha}\xi_i \in k[\sigma, \xi_1, \xi_2, \ldots, \xi_n]$; hence, (4.2) is a system of polynomial equations. Let

$$S = \left\{\sigma \xi^{\operatorname{mdeg}(m_\alpha)} - c_{i,\alpha}\xi_i \;\mid\; m_\alpha \in \operatorname{supp}(P),\ i = 1, 2, \ldots, n\right\},$$

and let $I = \langle S \rangle$. Thus, our problem of finding scaling factors σ and ξ_i so that parameters are eliminated is equivalent to determining $V_{\mathbb{R}}(I)$. If $V_{\mathbb{R}}(I) \neq \emptyset$, then we can eliminate all parameters by choosing $(\sigma, \xi_1, \xi_2, \ldots, \xi_n) \in V_{\mathbb{R}}(I)$ as scaling factors. If $V_{\mathbb{R}}(I) = \emptyset$ we can remove one of the equations of (4.2), $\sigma \xi^{\operatorname{mdeg}(m_{\alpha_1})} - c_{1,\alpha_1}\xi_1 = 0$ say, and then determine

$$V_{\mathbb{R}}\left(\left\langle S \backslash \left\{\sigma \xi^{\operatorname{mdeg}(m_{\alpha_1})} - c_{1,\alpha_1}\xi_1\right\}\right\rangle\right).$$

If this is still empty, remove another equation and try again, et cetera. Assume that after $j+1$ steps, the variety obtained is non-empty. Then at least j of the coefficients of the monomials in the support of the remaining polynomials are 1. Rename the remaining coefficients $\alpha_1, \alpha_2, \ldots, \alpha_j$. Then the parameters of the system are $\alpha_1, \alpha_2, \ldots, \alpha_j$, i.e. the number of parameters is j. If j is smaller than the number of parameters we started with, we have reduced the number of parameters. Thus, it is not guaranteed that the algorithm will reduce the number of parameters. However, it works in some cases, as the following example demonstrates.

Example 4.0.1. Consider the reduced system in Example 2.1.1, i.e.

$$\left\{\begin{array}{lcrcrcr} \dot{S} & = & -\lambda a S & + & \lambda S C & + & \mu C \\ \dot{C} & = & \lambda a S & - & \lambda S C & - & (\mu+\kappa)C \end{array}\right.$$

Let $X_1 = S$ and $X_2 = C$ and let x_i, ξ_i, τ and σ be defined as above. This gives

$$\begin{cases} p_1(\xi_1,\xi_2,\sigma) &= \sigma + \lambda a &= 0 \\ p_2(\xi_1,\xi_2,\sigma) &= -\xi_1\sigma + \lambda &= 0 \\ p_3(\xi_1,\xi_2,\sigma) &= \mu\xi_1 - \xi_2\sigma &= 0 \\ p_4(\xi_1,\xi_2,\sigma) &= -\xi_1\sigma + \lambda a\xi_2 &= 0 \\ p_5(\xi_1,\xi_2,\sigma) &= -\xi_1\sigma - \lambda &= 0 \\ p_6(\xi_1,\xi_2,\sigma) &= \sigma + \mu + \kappa &= 0 \end{cases}$$

It turns out that

$$\langle p_1, p_2, \dots, p_6\rangle = \mathbb{R}[\xi_1, \xi_2, \sigma]$$

and

$$\langle \{p_i \mid 1 \le i \le 6 \text{ and } i \ne j\}\rangle = \mathbb{R}[\xi_1, \xi_2, \sigma]$$

for all $j \in \{1, 2, \dots, 6\}$. But

$$J = \langle p_1, p_2, p_4, p_5\rangle$$

has the Gröbner basis (with respect to the monomial ordering $\xi_1 > \xi_2 > \sigma$

$$\{\sigma + a\lambda, a\xi_2 + 1, a\xi_1 + 1\},$$

which does not generate the whole ring $\mathbb{R}[\xi_1, \xi_2, \sigma]$. Thus, $V(J) \neq \emptyset$. Since the Gröbner basis has a strongly triangular form, the variety is ifinite. It is easily found, since each polynomial in the Gröbner basis is already univariate and linear: we see that $V(J) = \left\{\left(-a^{-1}, -a^{-1}, -a\lambda\right)\right\}$ (where the coordinates are in the order (ξ_1, ξ_2, σ). Let $s = x_1$ and $c = x_2$. Let

$$\alpha_1 = \frac{\mu}{-a\lambda}, \qquad \alpha_2 = \frac{\mu + \kappa}{-a\lambda}.$$

Then

$$\begin{cases} \dfrac{ds}{d\tau} = s - sc + \alpha_1 c \\ \dfrac{dc}{d\tau} = -s - sc - \alpha_2 c \end{cases}.$$

Thus, we have reduced the number of parameters from four (κ, λ, μ and a) to two: α_1 and α_2 . ◇

5 Computing the number of steady states

5.1 The trace formula

Consider the system $\dot{x}_i = p_i(x), i = 1, 2, \dots, n$, where $p_i \in \mathbb{R}[x]$. Let $I = \langle p_1, p_2, \dots, p_n\rangle$. As we have seen before, the set of steady states is given by $V_{\mathbb{R}}(I)$.

Assume that $|V_{\mathbb{R}}(I)| < \infty$. It turns out that we can determine the exact number of steady states without actually finding them. This follows from a result in [12], which we will present in this section.

While [12] proves that the result holds for polynomials over any subfield of a so called real closed field, we will — to avoid having to introduce the notion of a real closed field — assume that the polynomials are over $\mathbb{R}$ (which is a subfield of itself, and $\mathbb{R}$ is a real closed field).

Convention *Let V be an inner product space. Then $\langle x, y\rangle$ will denote the inner product of x and y, for $x, y \in V$.*

Remark *This notation clashes with our notation for an ideal generated by two elements (i.e. $\langle \cdot, \cdot\rangle$). However, in each particular case, the intended meaning should be clear from context.*

Now we define bilinear forms on real inner product spaces.

Definition 5.1.1. *Let V be a real inner product space over a field k. Let $B : V \times V \to k$ be a map which satisfies*

$$B(\alpha_1 v_1 + \alpha_2 v_2,\ \beta_1 w_1 + \beta_2 w_2) = \sum_{i=1}^{2}\sum_{j=1}^{2} \alpha_i \beta_j B(v_i, w_j);$$

i.e. it is linear in both arguments. Then we say that B is a bilinear form on V.

If $B(v, w) = B(w, v)$, we say that B is a symmetric bilinear form.

It turns out that, in finite-dimensional inner product spaces, all bilinear forms are of a certain form.

Proposition 5.1.2 ([8, Theorem 8.1])**.** *Let B be a bilinear form on a finite-dimensional inner product space V. Then there is a unique linear transformation $T : V \to V$ such that $B(v, w) = \langle Tv, w\rangle$ for all $v, w \in V$.*

Definition 5.1.3. *Let B be a bilinear form in V and let $\mathcal{B}$ be a basis of V. The unique linear transformation T_B such that $B(v, w) = \langle T_B v, w\rangle$ for all $v, w \in V$ will be called the linear transformation associated to B.*

It follows from Proposition 5.1.2 that, once a basis of V has been chosen, a bilinear form can be represented by a matrix.

Convention *In this section, we use $[v]_{\mathcal{B}}$ to denote the coordinate vector of $v \in V$ in basis $\mathcal{B}$, and $[T]_{\mathcal{B}}$ to denote the matrix of the linear transformation $T : V \to V$ in basis $\mathcal{B}$.*

Corollary 5.1.4. *Let B be a bilinear form on, and $\mathcal{B}$ a basis of, V. Then $B(x, y) = [x]_{\mathcal{B}}^T \, {[T_B]_{\mathcal{B}}}^T \, [y]_{\mathcal{B}}$ for all $x, y \in V$. If B is symmetric, then so is $[T_B]_{\mathcal{B}}$.*

Proof. Let $\mathcal{B} = \{v_1, v_2, \ldots, v_n\}$. Let $x = \sum_{i=1}^n c_i v_i$ and $y = \sum_{i=1}^n d_i v_i$. Then

$$\begin{aligned}
&B(x,y)\\
&= \langle \sum_{i=1}^n c_i T_B v_i, \sum_{i=1}^n d_i v_i \rangle\\
&= \sum_{i=1}^n \sum_{j=1}^n c_i d_j \langle T_B v_i, v_j \rangle\\
&= \begin{pmatrix} c_1 & c_2 & \ldots & c_n \end{pmatrix} \begin{pmatrix} \langle T_B v_1, v_1 \rangle & \langle T_B v_2, v_1 \rangle & \ldots & \langle T_B v_n, v_1 \rangle \\ \langle T_B v_1, v_2 \rangle & \langle T_B v_2, v_2 \rangle & \ldots & \langle T_B v_n, v_2 \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle T_B v_1, v_n \rangle & \langle T_B v_2, v_n \rangle & \ldots & \langle T_B v_n, v_n \rangle \end{pmatrix} \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_n \end{pmatrix}.
\end{aligned}$$

Let

$$\tilde{T} = \left(\langle T_B v_i, v_j \rangle\right)_{\substack{1 \le i \le n \\ 1 \le j \le n}}$$

and let $T = [T_B]_{\mathcal{B}}$. Note that $[v_i]_{\mathcal{B}}$ is a column vector with 1 in the i:th row and 0 in all other rows. Thus,

$$\langle T_B v_i, v_j \rangle = [v_i]_{\mathcal{B}}^T \, [T_B]_{\mathcal{B}}{}^T \, [v_j]_{\mathcal{B}} = t_{ji},$$

where t_{ji} denotes the element in the j:th row and the i:th column of T. Thus, $\tilde{T} = T^T$, i.e.

$$\left(\langle T_B v_i, v_j \rangle\right)_{\substack{1 \le i \le n \\ 1 \le j \le n}} = [T_B]_{\mathcal{B}}{}^T.$$

Since

$$[x]_{\mathcal{B}} = \begin{pmatrix} c_1 & c_2 & \ldots & c_n \end{pmatrix}^T$$

and

$$[y]_{\mathcal{B}} = \begin{pmatrix} d_1 & d_2 & \ldots & d_n \end{pmatrix}^T,$$

this gives $B(x,y) = [x]_{\mathcal{B}}^T \, [T_B]_{\mathcal{B}}{}^T \, [y]_{\mathcal{B}}$.

Note that $B(v_i, v_j) = \langle T v_i, v_j \rangle$. If B is symmetric, then $\langle T v_j, v_i \rangle = B(v_j, v_i) = B(v_i, v_j) = \langle T v_i, v_j \rangle$, so $\tilde{T}$ is symmetric; hence, $[T_B]_{\mathcal{B}}$ is symmetric. □

Definition 5.1.5. *The matrix* $\left([T_B]_{\mathcal{B}}\right)^T$ *in Corollary 5.1.4 will be called the matrix representation of* B *in basis* $\mathcal{B}$*.*

Note that

$$B(x,y) = \left([x]_{\mathcal{B}_1}\right)^T \left([T_B]_{\mathcal{B}_1}\right)^T [y]_{\mathcal{B}_1} = \left([x]_{\mathcal{B}_2}\right)^T P^T \left([T_B]_{\mathcal{B}_1}\right)^T P \, [y]_{\mathcal{B}_2}.$$

Thus, $\left([B]_{\mathcal{B}_2}\right)^T = P^T [B]_{\mathcal{B}_1} P$. [8, chapter 8.2.1]

We recall the following notion from elementary linear algebra.

Definition 5.1.6. *A matrix* A *is orthonormal if and only if* $A^T A$ *is the identity matrix.*

Note that a matrix A is orthonormal if and only if $A^{-1} = A^T$. We also recall, without proof, the following property of real and symmetric matrices.

Lemma 5.1.7. *Let A be a real and symmetric matrix. Then there exists an orthonormal matrix U such that $U^{-1}AU$ is a diagonal matrix (in particular, then, A is diagonalizble).*

Part (i) of the following proposition is called *Sylvester's law of inertia.*

Proposition 5.1.8 ([8, Theorem 8.10])**.** *Let B be a symmetric bilinear form on a real inner product space V. Then*

(i) there is a unique (up to reordering) basis $\mathcal{B}$ of V such that

$$B(x,y) = [x]_{\mathcal{B}}^T\, D\, [y]_{\mathcal{B}}\,,$$

where D is a diagonal matrix in which all diagonal elements belong to the set $\{-1,1,0\}$, and

(ii) the number of 1:s on the diagonal of D equals the number of positive eigenvalues (counted with algebraic multiplicity) and the number of -1:s the number of negative eigenvalues (counted with algebraic multiplicity) of B.

Proof. $[T_B]_{\mathcal{B}}$ is symmetric and real, so by Lemma 5.1.7, it is diagonalizable. Moreover, such a diagonalization is unique (up to reordering of the diagonal elements, which corresponds to reordering of the basis vectors). Thus, there is an orthonormal matrix U (i.e. $U^TU = I$, so $U^{-1} = U^T$) such that $U^T[T_B]_{\mathcal{B}}U = D_1$ for some diagonal matrix D_1, with the diagonal elements $\lambda_1, \lambda_2, \ldots, \lambda_n$ being eigenvalues of $[T_B]_{\mathcal{B}}$. Since we can reorder the diagonal elements of D_1 by reordering the basis vectors, we can without loss of generality assume that

$$\lambda_i \begin{cases} > 0, & \text{if } 1 \le i \le k \\ < 0, & \text{if } k+1 \le i \le m \\ = 0, & \text{if } m+1 \le i \le n \end{cases}.$$

Let $\mathcal{B}'$ be the basis with $[x]_{\mathcal{B}} = U\,[x]_{\mathcal{B}'}$. Let W be the diagonal matrix with diagonal elements $|\lambda_1|^{-1/2}, |\lambda_2|^{-1/2}, \ldots, |\lambda_n|^{-1/2}$. Let $\mathcal{B}''$ be the basis with $[x]_{\mathcal{B}'} = W\,[x]_{\mathcal{B}''}$. Then

$$\begin{aligned} B(x,y) &= [x]_{\mathcal{B}}^T\,[T_B]_{\mathcal{B}}\,[y]_{\mathcal{B}} \\ &= [x]_{\mathcal{B}'}^T\,U^T[T_B]_{\mathcal{B}}U\,[y]_{\mathcal{B}'} \\ &= [x]_{\mathcal{B}'}^T\,D_1\,[y]_{\mathcal{B}'} \\ &= [x]_{\mathcal{B}''}^T\,WD_1W\,[y]_{\mathcal{B}'} \end{aligned}$$

where we have used that $[T_B]_{\mathcal{B}}$ and W are symmetric. Note that WD_1W is a diagonal matrix D with

$$d_{ii} = \begin{cases} 1, & \text{if } 1 \le i \le k \\ -1, & \text{if } k+1 \le i \le m \\ 0, & \text{if } m+1 \le i \le n \end{cases}.$$

□

Recall the following notion from elementary linear algebra: two square matrices B and A, of the same size, are called similar if there is an invertible matrix P such that $B = P^{-1}AP$. Note that the matrix D in Proposition 5.1.8 is not neccesarily similar to $[T_B]_{\mathcal{B}}$: we only know that $D = (WU)^T [T_B]_{\mathcal{B}} (WU)$, and since WU is not neccessarily orthonormal it does not follow that D and $[T_B]_{\mathcal{B}}$ are similar.

Definition 5.1.9. *Let B be a symmetric bilinear form. Let D_B be the diagonal matrix in Proposition 5.1.8. Let n_+ be the number of 1:s, and n_- the number of -1:s, on the diagonal of D_B. The signature of B is defined*

$$\operatorname{sign}(B) = n_+ - n_-$$

Equivalently, by Proposition 5.1.8,

$$\operatorname{sign}(B) = \sum_{\lambda > 0} \mathrm{a.\,m.}\,(\chi, \lambda) - \sum_{\lambda < 0} \mathrm{a.\,m.}\,(\chi, \lambda)$$

where χ is the characteristic polynomial of $[T_B]_{\mathcal{B}}$ and $\mathrm{a.\,m.}\,(\chi, \lambda)$ is the multiplicity of λ as a root of χ.

Remark *The multiplicity of λ as a root of χ is called the algebraic multiplicity of λ as an eigenvalue of $[T_B]_{\mathcal{B}}$; hence, the notation $\mathrm{a.\,m.}\,(\cdot, \cdot)$.*

Definition 5.1.10. *Let V be a real inner product space over a field k. We say that $Q : V \times V \to k$ is a quadratic form, if there is a symmetric bilinear form B on V such that $Q(v) = B(v, v)$ for all $v \in V$.*

The quadratic form mentioned in the beginning of this section is the trace of a certain linear transformation. We introduce the quadratic form in three steps.

Definition 5.1.11. *Let $I \subset k[x]$ be an ideal. Let*

$$\begin{array}{rccc} T_p : & k[x]/I & \to & k[x]/I \\ & f + I & \mapsto & pf + I \end{array} .$$

Then T_p is called the linear transformation induced by multiplication with p.

Definition 5.1.12. *Let T_p be the linear transformation induced by multiplication with p. Let $q \in k[x]$. Let*

$$\begin{array}{rccc} B_q : & k[x]/I \times k[x]/I & \to & k \\ & (p_1, p_2) & \mapsto & \operatorname{tr}(T_{qp_1p_2}) \end{array} .$$

Then B_q is called the bilinear form induced by q.

Definition 5.1.13. *The quadratic form $Q_q(p) = B_q(p, p)$ is called the quadratic form induced by q.*

For proofs that T_p is a linear transformation and B_q is a symmetric linear form, see Appendix A.

Now we can formulate the result from [12].

Proposition 5.1.14 (cf. [12, Theorem 2.1]). *Let $q \in \mathbb{R}[x]$. Let I be an ideal such that $V(I)$ is finite. Let Q_q be the quadratic form on $\mathbb{R}[x]/I$ induced by q. Let $V_k^{q,R}(I) = \{x \in V_k(I) \mid q(x)\ R\ 0\}$, where $k \in \{\mathbb{R}, \mathbb{C}\}$ and $R \in \{>, <, \neq\}$. Then*

$$\begin{aligned} \operatorname{sign}(Q_q) &= \left|V_{\mathbb{R}}^{q,>}(I)\right| - \left|V_{\mathbb{R}}^{q,<}(I)\right|, \text{ and} \\ \operatorname{rank}(Q_q) &= \left|V_{\mathbb{C}}^{q,\neq}(I)\right| \end{aligned}$$

Corollary 5.1.15.

$$|V_{\mathbb{R}}(I)| = \operatorname{sign}(Q_1)$$

Proof. This follows from $V_{\mathbb{R}}^{1,<}(I) = \emptyset$ and $V_{\mathbb{R}}^{1,>}(I) = V_{\mathbb{R}}(I)$. □

To use Proposition 5.1.14, we need a basis of $\mathbb{R}[x]/I$. Fortunately, such a basis can be found using Gröbner bases, by Proposition 3.8.1 (recall that $\ell(I) = \langle\{\operatorname{lm}(g) \mid g \in G\}\rangle$ if G is a Gröbner basis of I). From that proposition, the following follows.

Corollary 5.1.16. *Let $\dot{x}_i = p_i(x)$ and let $P = \{p_1, p_2, \ldots, p_n\}$. Fix an enumeration μ of P and a monomial ordering $<$. Let $C_{<,\mu}$ be the corresponding matrix form. Assume that $C_{<,\mu}$ has full rank. Let $j_k = \min\{i \mid c_{ki} \neq 0\}$ for $k = 1, 2, \ldots, n$. Then $\operatorname{mon}(k[x]) \setminus \{m_{j_k} \mid k \in \{1, 2, \ldots, n\}\}$ is a basis for $k[x]/\langle P\rangle$.*

Proof. Note that $\operatorname{lm}(p_k) = m_{j_k}$, by construction. The conclusion now follows immediately from Proposition 3.8.1. □

Definition 5.1.17. *Let $c = (c_1, c_2, \ldots, c_n) \in \mathbb{R}^n$. Then*

$$\mathrm{s.\,c.}(c) = |\{j \in \mathbb{N} \mid c_j c_{j+1} < 0\}|$$

is the number of sign changes in this n-tuple.

Let $f = \sum_{i=0}^{n} c_i t^i$, with $c_i \in \mathbb{R}$. Then $\mathrm{c}_f = (c_0, c_1, \ldots, c_n)$ is called the sequence of coefficients of f.

For convenience, we define $\mathrm{s.\,c.}(f) = \mathrm{s.\,c.}(\mathrm{c}_f)$.

Definition 5.1.18. *Let $f \in k[t]$ be a univariate polynomial, such that all its roots are real. The number of positive roots of f (counted with multiplicity) is denoted $\mathrm{n.\,r.}(f, +)$. The number of negative roots of f (counted with multiplicity) is denoted $\mathrm{n.\,r.}(f, -)$.*

It turns out that, if all roots of a polynomial are real, we can compute the number of positive roots by counting the number of sign changes in the sequence of coefficients of f. To prove this, we will use the following lemma.

Lemma 5.1.19. *Let $f \in k[t]$ be a univariate polynomial, such that all its roots are real and have multiplicity one. Then f' has precisely one root between each pair of consecutive roots of f, and f has precisely one root between each consecutive pair of roots of f'.*

Proof. Let $\deg(f) = n$. Let a and b be consecutive roots of f, i.e. $f(x) \neq 0$ on $a < x < b$. From the mean value theorem, it follows that there is a ξ such that $a < \xi < b$ and $f'(\xi) = 0$. Thus, between any pair of consecutive roots of f, there is at least one root of f'. Since f has n roots, the number of roots of

f' accounted for in this way is $n-1$. But f' is a polynomial of degree $n-1$, so this accounts for all its roots. Thus, we must have the alternating pattern

$$t_1, \tilde{t}_1, t_2, \tilde{t}_2, t_3, \ldots, t_{n-1}, \tilde{t}_n, t_n,$$

where the t_i are the roots of f and $\tilde{t}_i$ the roots of f'. This proves the statements in the lemma. □

Proposition 5.1.20. *Let $f \in k[t]$ be a univariate polynomial such that all its roots are real. Then* $\mathrm{n.\,r.}\,(f,+) = \mathrm{s.\,c.}\,(f)$.

Proof. Let $f(t) = \sum_{i=0}^{n} c_i t^i$. If $c_0 = 0$, then $f(0) = 0$, so $f(t) = t^m g(t)$ for some polynomial g. We can choose m so that $g(t)$ is not divisible by any power of t, which is equivalent to $g(0) \neq 0$. Since 0 is not positive, the positive roots of f and g are the same. Thus, we can, without loss of generality, assume that $f(0) \neq 0$, i.e. $c_0 = 0$.

First, assume that all roots of f has multiplicity one. Note that $f(0) = c_0$ and $f'(0) = c_1$. Either $c_0c_1 > 0$ or $c_0c_1 < 0$.

If $c_0c_1 > 0$, then $f(0)$ and $f'(0)$ have the same sign. Assume that $f(0)$ and $f'(0)$ are positive. Let t_1 be the first positive root of f, i.e. $f(t_1) = 0$ and $f(t) \neq 0$ on $0 \leq t_1$. Since f is positive and increasing at $t = 0$, but f is zero at t_1, there must be $\tilde{t}$ such that $f'(\tilde{t}) < 0$, so there must be a $\tilde{t}_1$ such that $f'(\tilde{t}_1) = 0$ for some $\tilde{t}_1$ such that $0 < \tilde{t}_1 < t_1$. Assume that $\mathrm{n.\,r.}\,(f,+)$ is n_+. By Lemma 5.1.19, each pair of consecutive roots of f contributes one root of f'; thus, this contributes $n_+ - 1$ roots of f'. Since there is one positive root of f' before the first positive root of f, we have $\mathrm{n.\,r.}\,(f,+) = \mathrm{n.\,r.}\,(f',+)$. The case $f(0), f'(0) < 0$ follows from symmetry.

If $c_0c_1 < 0$, then $f(0)$ and $f'(0)$ have opposite signs. Assume that $f(0) > 0$ and $f'(0) < 0$. Since f is positive and decreasing at $t = 0$, Lemma 5.1.19 implies that there must be a t_{-1} and $\tilde{t}_{-1}$ such that t_{-1} is a root of f and $\tilde{t}_{-1}$ a root of f', with

- $t_{-1} < \tilde{t}_{-1}$,
- no root of f in the interval $t_1 < t < 0$, and
- no root of f' in the interval $\tilde{t}_1 < t < 0$.

By Lemma 5.1.19, there is a $t_1 > 0$ such that neither f nor f' has any root in the interval $0 < t < t_1$ (informally: the first positive root of f "comes before" the first positive root of f'). It follows Lemma 5.1.19 that

$$\mathrm{n.\,r.}\,(f,+) = \mathrm{n.\,r.}\,(f',+) + 1.$$

The case $f(0) < 0$ and $f'(0) > 0$ follows from symmetry.

Note that $f'(t) = \sum_{k=1}^{n} kc_k t^{k-1}$. We have $\mathrm{c}_f = (c_0, c_1, c_2, \ldots, c_n)$ and $\mathrm{c}_{f'} = (c_1, 2c_2, 3c_3, \ldots, nc_n)$. If $c_0c_1 > 0$, then

$$\mathrm{s.\,c.}\,(f) = \mathrm{s.\,c.}\,(f')\,.$$

If $c_0c_1 < 0$, then

$$\mathrm{s.\,c.}\,(f) = \mathrm{s.\,c.}\,(f') + 1.$$

Now we can prove the statement by induction over the degree of f. If

$$f(t) = t - \alpha,$$

then $\mathrm{c}_f = (-\alpha, 1)$.

If $\alpha > 0$, then $\mathrm{n.\,r.}\,(f, +) = 1$, and $\mathrm{s.\,c.}\,(f) = 1$. If $\alpha < 0$, then $\mathrm{n.\,r.}\,(f, +) = 0$, and $\mathrm{s.\,c.}\,(f) = 0$. Thus, the statement holds for polynomials of degree 1.

Assume that the statement holds for polynomials of degree k. Let $\deg(f) = k + 1$. Then $\deg(f') = k$, so

$$\mathrm{n.\,r.}\,(f', +) = \mathrm{s.\,c.}\,(f'),$$

by the induction hypothesis. If $c_0c_1 > 0$, then

$$\begin{aligned} \mathrm{n.\,r.}\,(f, +) &= \mathrm{n.\,r.}\,(f', +) \\ \mathrm{s.\,c.}\,(f) &= \mathrm{s.\,c.}\,(f'); \end{aligned}$$

hence,

$$\mathrm{n.\,r.}\,(f, +) = \mathrm{s.\,c.}\,(f).$$

If $c_0c_1 < 0$, then

$$\begin{aligned} \mathrm{n.\,r.}\,(f, +) &= \mathrm{n.\,r.}\,(f', +) + 1 \\ \mathrm{s.\,c.}\,(f) &= \mathrm{s.\,c.}\,(f') + 1; \end{aligned}$$

hence,

$$\mathrm{n.\,r.}\,(f, +) = \mathrm{s.\,c.}\,(f).$$

It remains to show the statement for f in which not all roots have multiplicity one. Let $\hat{t}$ be a root of f which has multiplicity $m > 1$. Then $\hat{t}$ is a root of f' of multiplicity $m - 1$. Hence, a root of f of multiplicity m contributes $m - 1$ roots $\mathrm{n.\,r.}\,(f', +)$ (since we count with multiplicity). Thus, the statement holds for this case as well. □

Remark *The proposition is a stronger version of the result known as "Descartes' rule of signs". For polynomials which has complex roots, we can only say* $\mathrm{n.\,r.}\,(f, +) \equiv_2 \mathrm{s.\,c.}\,(f)$.

Corollary 5.1.21. *Let $f \in k[t]$ be a univariate polynomial such that all its roots are real. Let $(f^-)(t) = f(-t)$. Then* $\mathrm{n.\,r.}\,(f, -) = \mathrm{s.\,c.}\,(f^-)$.

Recall that the characteristic polynomial χ_A of a matrix A is defined $\chi_A(t) = \det(\lambda I - A)$.

Corollary 5.1.22. *Let $T : V \to V$ be a linear transformation, such that all eigenvalues of T are real. Let $\chi_T(t)$ be the characteristic polynomial of T. Then the number of positive eigenvalues of T is equal to* $\mathrm{s.\,c.}\,(\chi_T)$ *and the number of negative eigenvalues of T is equal to* $\mathrm{s.\,c.}\,((\chi_T)^-)$.

Example 5.1.23. We continue Example 3.8.6. Since $\{g_1, g_2, g_3\}$ is a Gröbner basis of I, we have $\ell(I) = \langle x_1^2, x_2, x_3^6 \rangle$. Thus,

$$\mathcal{B} = \left\{1,\ x_3,\ x_3^2,\ x_3^3,\ x_3^4,\ x_3^5,\ x_1,\ x_1x_3,\ x_1x_3^2,\ x_1x_3^3,\ x_1x_3^4,\ x_1x_3^5\right\}$$

(where each term actually denotes its equivalence class in $\mathbb{R}[x_1, x_2, x_3]/I$) is a basis for $\mathbb{R}[x_1, x_2, x_3]/I$. We need to compute $\operatorname{tr}\left(T_{x_1^i x_3^j}\right)$ for $i = 0, 1$ and $j = 0, 1, 2, 3, 4, 5$. Since $T_{x_1^i x_3^j} = T_{x_1}^i \circ T_{x_3}^j$, it is sufficient to compute T_{x_1} and T_{x_3}. We get

$$[T_{x_3}]_{\mathcal{B}} = \left(\begin{array}{cccccc|cccccc}
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0
\end{array}\right)$$

and

$$[T_{x_1}]_{\mathcal{B}} = \left(\begin{array}{cccccc|cccccc}
0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\
\hline
1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\
0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0
\end{array}\right).$$

After some computation, we get

$$[Q_1]_{\mathcal{B}} = \left(\begin{array}{rrrrrr|rrrrrr}
12 & 0 & 0 & 0 & 0 & 0 & 0 & 6 & 0 & 0 & 6 & 0 \\
0 & 0 & 0 & 0 & 0 & 12 & 6 & 0 & 0 & 6 & 0 & 0 \\
0 & 0 & 0 & 0 & 12 & 0 & 0 & 0 & 6 & 0 & 0 & 6 \\
0 & 0 & 0 & 12 & 0 & 0 & 0 & 6 & 0 & 0 & 6 & 0 \\
0 & 0 & 12 & 0 & 0 & 0 & 6 & 0 & 0 & 6 & 0 & 0 \\
0 & 12 & 0 & 0 & 0 & 0 & 0 & 0 & 6 & 0 & 0 & 6 \\
\hline
0 & 6 & 0 & 0 & 6 & 0 & -12 & 0 & 12 & 0 & 0 & 12 \\
6 & 0 & 0 & 6 & 0 & 0 & 0 & 12 & 0 & 0 & 12 & -12 \\
0 & 0 & 6 & 0 & 0 & 6 & 12 & 0 & 0 & 12 & -12 & 0 \\
0 & 6 & 0 & 0 & 6 & 0 & 0 & 0 & 12 & -12 & 0 & 12 \\
6 & 0 & 0 & 6 & 0 & 0 & 0 & 12 & -12 & 0 & 12 & 0 \\
0 & 0 & 6 & 0 & 0 & 6 & 12 & -12 & 0 & 12 & 0 & 0
\end{array}\right).$$

The characteristic polynomial of Q_1 is

$$\begin{aligned}\chi_{Q_1}(t) = t^{12} &- 24t^{11} - 1872t^{10} + 44928t^9 + 746496t^8 - 17915904t^7 \\ &- 119439360t^6 + 2866544640t^5 + 8169652224t^4 - 196071653376t^3 \\ &- 185752092672t^2 + 4458050224128t.\end{aligned}$$

Let $(\chi_{Q_1})^-(t) = \chi_{Q_1}(-t)$. The number of sign changes in the sequence of coefficients of χ_{Q_1} is 6, while the number of sign changes in $(\chi_{Q_1})^-$ is 5. Thus, the signature of Q_1 is 1, so the system has one steady state. This is in agreement with Example 3.8.6. ◇

5.2 Proof of the trace formula

The proof will follow the structure of the proof given in [12, p. 209–210]. The idea of the proof is to decompose the ring $\mathbb{C}[x]/I$, which is also a vector space over $\mathbb{C}$, and show that each component is invariant under the transformation induced by p, where $p \in \mathbb{R}[x]$ is arbitrary. This is then used to give an explicit formula for B_p, from which the statement will follow.

For the decomposition part of the proof, we will show how it follows from the theory of so called Artinian rings. The authors of [12] in their article remark that this can be done, but they show the decomposition in a different way.

5.2.1 Decomposition

Our first goal is to prove the following lemma.

Lemma 5.2.1 ([12, chapter 2])**.** *Let $V(I) = \{\alpha_i \mid 1 \leq i \leq k\}$. Then*

$$\mathbb{C}[x]/I\mathbb{C}[x] \cong \prod_{i=1}^{k} (\mathbb{C}[x]/I\mathbb{C}[x])_{m_{\alpha_i}/I\mathbb{C}[x]}.$$

The notation will be explained as we go along. First, let us introduce the notion of an Artinian ring.

Definition 5.2.2 ([1, chapter 6])**.** *A chain of ideals $(I_j)_{j=1}^{\infty}$ such that $I_j \subsetneq I_{j+1}$ for all j is called a descending chain of ideals. If there is a $k \in \mathbb{N}$ such that $I_{k+i} = I_k$ for all $i \in \mathbb{N}$, we say that the chain satisfies the descending chain condition.*

Let R be a ring such that every descending chain of ideals $(I_j)_{j=1}^{\infty}$ satisfies the descending chain condition. Then we say that R is Artinian.

Proposition 5.2.3 (cf. [1, exercise 3 in chapter 8])**.** *Let R be a ring which is also a finite-dimensional vector space over a field k. Then R is Artinian.*

Proof. Let $(I_j)_{j=1}^{n}$ be a decreasing chain of ideals. Each ideal is closed under addition, so each ideal is also a k-subspace of R. Moreover, I_{j+1} is a k-subspace of I_j. Let $\dim_k R = n$. Then $\dim_k I_1 < n$. We also have $\dim_k I_{j+1} < \dim_k I_j$. Assume $I_j \subsetneq I_{j+1}$. Then $(\dim_k I_j)_{j=1}^{\infty}$ is a strictly decreasing sequence of positive numbers. But $\dim_k I_j \geq 0$ for every j, so this is impossible. Hence, there is an m such that $I_{m+i} = I_m$ for all $i \geq 0$. Hence, R is Artinian. □

Let I be an ideal of $\mathbb{R}[x]$. Let $\{p_1, p_2, \dots, p_k\}$ be a set of generators of I. This means $I = \left\{\sum_{i=1}^{k} q_i p_i \mid q_i \in \mathbb{R}[x]\right\}$. Since $\mathbb{R}[x] \subset \mathbb{C}[x]$, this ideal induces an ideal $I\mathbb{C} = \left\{\sum_{i=1}^{k} q_i p_i \mid q_i \in \mathbb{C}[x]\right\}$ of $\mathbb{C}[x]$.

Definition 5.2.4 (cf. [12, chapter 2]). *Let $k \subset K$ be fields. Let $I = \langle p_1, p_2, \dots, p_k \rangle$ be an ideal of $k[x]$. The ideal $IK = \left\{\sum_{i=1}^{k} q_i p_i \mid q_i \in K[x]\right\}$ will be called the ideal of $K[x]$ induced by I.*

We have the following corollary of Proposition 5.2.3

Corollary 5.2.5. *Let $I \subset \mathbb{R}[x]$ be an ideal such that $V(I)$ is finite. Then $\mathbb{C}[x]/I\mathbb{C}[x]$ is Artinian.*

Proof. $I \subset \mathbb{R}[x] \subset \mathbb{C}[x]$, so $I\mathbb{C}[x]$ is an ideal of $\mathbb{C}[x]$. $V(I) = V(I\mathbb{C})$ is finite, so Lemma 3.8.2 implies that that $\mathbb{C}[x]/I\mathbb{C}[x]$ is finite-dimensional. By the proposition, this implies that $\mathbb{C}[x]/I\mathbb{C}[x]$ is Artinian. □

Recall the following notion from elementary abstract algebra.

Definition 5.2.6 ([1, chapter 1]). *Let R be a ring with an ideal I. Assume that there is no ideal $J \subset R$ such that $I \subsetneq J \subsetneq R$. Then we say that I is a maximal ideal.*

It turns out that Artinian rings have finitely many maximal ideals.

Proposition 5.2.7 ([1, Proposition 8.3]). *Let A be an Artinian ring. Then A has finitely many maximal ideals.*

Definition 5.2.8. *Let $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_n) \in \mathbb{C}^n$. Then*

$$m_\alpha = \langle \{x_i - \alpha_i \mid 1 \leq i \leq n\} \rangle \subset \mathbb{C}[x]$$

is the maximal ideal in $\mathbb{C}[x]$ corresponding to α.

For algebraically closed fields k, the maximal ideals of $k[x]$ have a certain form.

Proposition 5.2.9 ([14, Corollary 5.2]). *Let k be an algebraically closed field. Then m is a maximal ideal of $k[x]$ if and only if $m = \langle \{x_i - \alpha_i \mid 1 \leq i \leq n\} \rangle$ for some $\alpha \in k^n$.*

The following proposition gives the maximal ideals of $\mathbb{C}[x]/I\mathbb{C}[x]$.

Proposition 5.2.10. *Let $V(I\mathbb{C}[x]) = V(I) = \{\alpha_1, \alpha_2, \dots, \alpha_k\}$. Then*

$$m_{\alpha_i}/I\mathbb{C}[x],\ i = 1, 2, \dots, k$$

are the only maximal ideals in $\mathbb{C}[x]/I\mathbb{C}[x]$.

Proof. Let $M \supset I\mathbb{C}[x]$ be a maximal ideal of $\mathbb{C}[x]/I\mathbb{C}[x]$. Recall from elementary abstract algbra that for rings R with ideals I, there is a one-to-one-correspondence between ideals of R/I and ideals J of R such that $I \subset J$; more precisely, the ideal J corresponds to the ideal $J/I = \{j + I \mid j \in J\}$. In particular, maximal ideals of R which contain I corresponds to maximal ideals of R/I. Thus, we can assume

$$M = m/I\mathbb{C}[x]$$

for some maximal ideal m of $\mathbb{C}[x]$. Then

$$m = \langle \{x_i - \beta_i \mid 1 \leq i \leq n\} \rangle$$

for some $\beta = (\beta_1, \beta_2, \ldots, \beta_n)$, by Proposition 5.2.9. Take $f \in I\mathbb{C}[x]$. Since $I\mathbb{C}[x] \subset m$, this implies that $f(\beta) = 0$. Therefore, $\beta \in V(I)$. In other words, $\beta = \alpha_i$ for some $i \in \{1, 2, \ldots, k\}$, so $m = \langle \{x_j - \alpha_{ij} \mid 1 \leq j \leq n\} \rangle$, where $\alpha_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{in})$.

The correspondence between maximal ideals of $\mathbb{C}[x]/I\mathbb{C}[x]$ and maximal ideals of $\mathbb{C}[x]$ containing I together with Proposition 5.2.9 implies that $m_{\alpha_i}/I,\ i = 1, 2, \ldots, k$, are maximal ideals of $\mathbb{C}[x]/I\mathbb{C}[x]$. □

Artinian rings have a decomposition into a product of so called localizations at the maximal ideals of the ring. We introduce the notion of localizations of a ring in two steps.

Definition 5.2.11 ([1, chapter 3])**.** *Let R be a commutative ring. Let $S \subset R$ be a subset such that $1 \in S$ and $xy \in S$ for every $x, y \in S$. Then we say that S is a multiplicative subset of R.*

Let $p \subset R$ be a prime ideal. Let $S = R \backslash p$. Take $x, y \in S$, i.e. $x, y \notin p$. Then, by the definition of prime ideals, it follows that $xy \notin p$, i.e. $xy \in S$. Thus, S is a multiplicative subset of S. The multiplicative subsets of interest to us are precisely those which arise in this way.

Proposition 5.2.12 ([1, chapter 3])**.** *Let R be a ring with a multiplicative subset S. The quotient of the set $\{(r, s) \mid r \in R$ and $s \in S\}$ with the relation*

$$(r_1, s_1) \sim (r_2, s_2) \ \Leftrightarrow\ \exists t \in S:\ t(r_1 s_2 - r_2 s_1) = 0$$

is a ring under the operations

$$\begin{aligned}(r_1, s_1) + (r_2, s_2) &= (r_1 s_2 + r_2 s_1,\ s_1 s_2)\\ (r_1, s_1) \cdot (r_2, s_2) &= (r_1 r_2, s_1 s_2).\end{aligned}$$

Such a ring is called a ring of fractions and is denoted $S^{-1}R$.

Remark *The elements of $S^{-1}R$ are often denoted $\frac{r}{s}$.*

When $S = R \backslash p$ for a prime ideal p, the ring $S^{-1}R$ has a special name and a special notation is used.

Definition 5.2.13 ([14, chapter 6.4])**.** *Let R be a ring with a prime ideal p. Let $S = R \backslash p$. Then $R_p = S^{-1}R$ is called the localization of R at p.*

Let A be an Artinian ring, and let $m_i,\ i = 1, 2, \ldots, k$ be the maximal ideals of A. Recall from elementary abstract algebra that every maximal ideal is prime. Thus, we can localize at m_i.

Proposition 5.2.14 ([1, Theorem 8.7])**.** *Let A be an Artinian ring. Let $m_i,\ i = 1, 2 \ldots, k$, be its maximal ideals. Then $A = \prod_{i=1}^{k} A/m_i^r$ for some $r \in \mathbb{N}$.*

Let

$$\begin{array}{rccc} \phi_i: & A & \to & A_{m_i} \\ & a & \mapsto & \frac{a}{1} \end{array}.$$

Then $\phi_i(a) = 0$ if and only if there is an $s \in A \backslash m_i$ such that $sa = 0$ (where 0 is the additive identity in A). The following lemma, called *Nakayama's lemma*, will enable to show that $\ker \phi_i = m_i^r$.

Lemma 5.2.15 (cf. [14, Corollary 2 in chapter 2.8]). *Let $I \subset A$ be an ideal of a ring A with unique maximal ideal m. If $mI = I$, then $I = 0$.*

Since

$$A_{m_i} \supset m_i A_{m_i} \supset (m_i A_{m_i})^2 \supset \ldots$$

and A_{m_i} is Artinian, there is an $s_i \in \mathbb{N}$ such that $(m_i A_{m_i})^{s_i+1} = (m_i A_{m_i})^{s_i}$. Since

$$(m_i A_{m_i})^{s_i+1} = m_i (m_i A_{m_i})^{s_i},$$

if follows from Lemma 5.2.15 that $(m_i A_{m_i})^{s_i} = 0$ (where 0 is the additive identity in A_{m_i}). Let $r = \max \{s_i \mid i = 1, 2, \ldots, k\}$. Note that $m_i^{s_i} \supset m_i^r$ for all i. If $a \in m_i^r$, then $a \in m_i^{s_i}$. But since $m_i^{s_i} A_{m_i}$, it follows that $\phi_i(a) = 0$. Thus, $m_i^r \subset \ker \phi_i$. On the other hand, assume that $a \in \ker \phi_i$. Then there is an $s \in A \backslash m_i$ such that $sa = 0$ in A. Since $m_i^r A_{m_i} \subset m_i^{s_i} A_{m_i} = 0$, this implies that $sa \in m_i^r A_{m_i}$. Now we are almost finished, but we need one more concept and result.

Definition 5.2.16 ([1, chapter 4]). *Let I be an ideal of a ring A. Assume that the following implication holds:*

$$xy \in I \ \Rightarrow \left[x \in I \ \text{or} \ y^k \in I \ \text{for some} \ k \in \mathbb{N}\right].$$

Then we say that I is primary.

Lemma 5.2.17 ([1, Proposition 4.2]). *Let m be a maximal ideal of a ring A. Then m^k is primary, for every $k \in \mathbb{N}$.*

Since $m_i A_{m_i}$ is maximal, Lemma 5.2.17 implies that $(m_i A_{m_i})^r$ is primary. Since $s \in A \backslash m_i$ and $m_i A_{m_i} \supset (m_i A_{m_i})^k$ for every k, it follows that $a \in m_i^r A_{m_i}$. Thus, $\ker \phi_i \subset m_i^r$. We conclude that $\ker \phi_i = m_i^r$, so $A_{m_i} \cong A/m_i^r$.

Now Lemma 5.2.1 follows: take $A = \mathbb{C}[x]/I\mathbb{C}[x]$ and note that $m_i/I\mathbb{C}[x]$, $i = 1, 2, \ldots, k$, are the maximal ideals of $\mathbb{C}[x]/I\mathbb{C}[x]$, by Proposition 5.2.10.

5.2.2 Invariance

Lemma 5.2.18. *Let $V(I) = \{\alpha_1, \alpha_2, \ldots, \alpha_k\}$. Let T_p be the linear transformation $\mathbb{C}[x]/I\mathbb{C}[x] \to \mathbb{C}[x]/I\mathbb{C}[x]$ induced by p. Then*

(i) $(\mathbb{C}[x]/I\mathbb{C}[x][x])_{m_{\alpha_i}/I\mathbb{C}[x]}$ is invariant under T_p [12, Lemma 2.5];

(ii) there is a basis $\mathcal{B}_i$ of $(\mathbb{C}[x]/I\mathbb{C}[x][x])_{m_{\alpha_i}/I}$ such that $[T_p]_{\mathcal{B}_i}$ (where T_p is taken as restricted to $(\mathbb{C}[x]/I\mathbb{C}[x][x])_{m_{\alpha_i}/I}$) is upper triangular and in which all diagonal elements are equal to $p(\alpha_i)$ [12, Lemma 2.6 and the paragraph following it].

Proof. (i) The proof of this part follows the proof in the cited article. The elements of $(\mathbb{C}[x]/I\mathbb{C}[x])_{m_\alpha/I\mathbb{C}[x]}$ have the form

$$\frac{f+I\mathbb{C}[x]}{g+I\mathbb{C}[x]}$$

where $f+I\mathbb{C}[x],\ g+I\mathbb{C}[x]\in\mathbb{C}[x]/I\mathbb{C}[x]$ but $g+I\mathbb{C}[x]\notin m_\alpha/I\mathbb{C}[x]$. Then

$$T_p\left(\frac{f+I\mathbb{C}[x]}{g+I\mathbb{C}[x]}\right)=\frac{pf+I\mathbb{C}[x]}{g+I\mathbb{C}[x]}\in(\mathbb{C}[x]/I\mathbb{C}[x])_{m_\alpha/I\mathbb{C}[x]},$$

since $g+I\mathbb{C}[x]\notin m_\alpha/I\mathbb{C}[x]$.

(ii) Let $p\in\mathbb{C}[x]$. Set $p_i(x)=p(x)-p(\alpha_i)$ (the author got the idea to consider this polynomial from the proof of [12, Lemma 2.6]). Let

$$q_i(x)=\frac{1}{k}\sum_{j\neq i}^{k}\prod_{m=1}^{n}\frac{x_j-\alpha_{mj}}{\alpha_{ij}-\alpha_{mj}}$$

(the inspiration to use this polynomial came from the proof of [2, Proposition 4.91]). We see that

$$q_i(\alpha_j)=\begin{cases}1, j=i\\0, j\neq i\end{cases}.$$

Also, $p_i(\alpha_i)=0$. Hence, $q_ip_i(\alpha)=0$ for all $\alpha\in V(I)$, from which it follows that $q_ip_i\in\sqrt{I\mathbb{C}[x]}$, by Hilbert's Nullstellensatz [14, Theorem 5.6], i.e. there is a $k\in\mathbb{N}$ such that $(q_ip_i)^k\in I\mathbb{C}[x]$. We have

$$T_{p_i}^k\left(\frac{f+I\mathbb{C}[x]}{g+I\mathbb{C}[x]}\right)=\frac{p_i^kf+I\mathbb{C}[x]}{g+I\mathbb{C}[x]}.$$

But note that $q_i^k\notin m_{\alpha_i}$ and $\left(q_i^k+I\mathbb{C}[x]\right)\left(p_i^kf+I\mathbb{C}[x]\right)=q_i^kp_i^kf+I\mathbb{C}[x]=I\mathbb{C}[x]$, so

$$\frac{p_i^kf+I\mathbb{C}[x]}{g+I\mathbb{C}[x]}=0$$

in $(\mathbb{C}[x]/I\mathbb{C}[x][x])_{m_{\alpha_i}/I\mathbb{C}[x]}$, by the definition of ring of fractions. Hence, $T_{p_i}^k$ is zero, which implies that T_{p_i} is nilpotent, on $(\mathbb{C}[x]/I\mathbb{C}[x])_{m_{\alpha_i}/I\mathbb{C}[x]}$. Choose a basis $\mathcal{B}_i$ of $(\mathbb{C}[x]/I\mathbb{C}[x])_{m_{\alpha_i}/I\mathbb{C}[x]}$ such that $[T_{p_i}]_{\mathcal{B}_i}$ is upper triangular with a diagonal consisting only of zeros. Since $T_{p_i}=T_p-T_{p(\alpha_i)}$, we have

$$[T_{p_i}]_{\mathcal{B}_i}=[T_p]_{\mathcal{B}_i}-\left[T_{p(\alpha_i)}\right]_{\mathcal{B}_i}.$$

Furthermore, since $T_{p(\alpha_i)}$ is the linear transformation induced by multiplication of the constant polynomial $p(\alpha_i)$, the matrix $\left[T_{p(\alpha_i)}\right]_{\mathcal{B}_i}$ is a diagonal matrix with all diagonal elements equal to $p(\alpha_i)$. Since the diagonal elements of $[T_{p_i}]_{\mathcal{B}_i}$ are all equal to zero, this implies that all diagonal elements of $[T_p]_{\mathcal{B}_i}$ is also equal to $p(\alpha_i)$. Moreover, since $[T_{p_i}]_{\mathcal{B}_i}$ is upper triangular and $\left[T_{p(\alpha_i)}\right]_{\mathcal{B}_i}$ diagonal, we must have that $[T_p]_{\mathcal{B}_i}$ is also upper triangular. Thus, we have shown that there is a basis $\mathcal{B}_i$ of $(\mathbb{C}[x]/I\mathbb{C}[x])_{m_{\alpha_i}/I\mathbb{C}[x]}$ such that $[T_p]_{\mathcal{B}_i}$ is an upper triangular matrix with all diagonal elements equal to $p(\alpha_i)$.

□

5.2.3 Formula for the bilinear form

Let $\mathcal{B}_i$ be a basis of $(\mathbb{C}[x]/I\mathbb{C}[x])_{m_{\alpha_i}/I\mathbb{C}[x]}$, for $i = 1, 2, \ldots, k$, which satisfies part (ii) of Lemma 5.2.18. Taken together, they form a basis $\mathcal{B} = \{v_1, v_2, \ldots, v_s\}$ of $\mathbb{C}[x]/I\mathbb{C}[x]$. Let d_i be the dimension of $(\mathbb{C}[x]/I\mathbb{C}[x])_{m_{\alpha_i}/I\mathbb{C}[x]}$ as a vector space over $\mathbb{C}$. Then $[T_p]_{\mathcal{B}_i}$ is a square matrix of order d_i. Since each $(\mathbb{C}[x]/I\mathbb{C}[x])_{m_{\alpha_i}/I\mathbb{C}[x]}$ is invariant under T_p, by Lemma 5.2.18, the matrix $[T_p]_{\mathcal{B}}$ is a block diagonal matrix in which the i:th block is an upper triangular square matrix of order d_i with all diagonal elements equal to $p(\alpha_i)$.

This gives

$$\begin{aligned} B_q(p_1, p_2) &= \operatorname{tr}\left(T_{qp_1p_2}\right) \\ &= \operatorname{tr}\left([T_{qp_1p_2}]_{\mathcal{B}}\right) \\ &= \sum_{i=1}^{k} d_i q(\alpha_i) p_1(\alpha_i) p_2(\alpha_i); \end{aligned}$$

in particular,

$$\begin{aligned} B_q(v_i, v_j) &= \operatorname{tr}\left(T_{qp_1p_2}\right) \\ &= \operatorname{tr}\left([T_{qp_1p_2}]_{\mathcal{B}}\right) \\ &= \sum_{s=1}^{k} d_s q(\alpha_s) v_i(\alpha_s) v_j(\alpha_s). \end{aligned}$$

The following lemma implies that this formula for B_q remains valid if restricted to $\mathbb{R}[x]/I$.

Lemma 5.2.19 ([2, Lemma 4.86])**.** $\mathbb{R}[x]/I \subset \mathbb{C}[x]/I\mathbb{C}[x]$

Let $V(I) = \{\alpha_1, \alpha_2, \ldots, \alpha_k\}$ and let $\mathcal{B} = \{v_1, v_2, \ldots, v_m\}$ be a basis of $\mathbb{R}[x]/I$. Let

$$V = \begin{pmatrix} v_1(\alpha_1) & v_1(\alpha_2) & \ldots & v_1(\alpha_k) \\ v_2(\alpha_1) & v_2(\alpha_2) & \ldots & v_2(\alpha_k) \\ \vdots & \vdots & \vdots & \vdots \\ v_m(\alpha_1) & v_m(\alpha_2) & \ldots & v_m(\alpha_k) \end{pmatrix}$$

and let

$$D = (d_{ij})_{\substack{1\le i\le n \\ 1\le j\le n}}$$

be the diagonal matrix of order k with $d_{ii} = d_i q(\alpha_i)$ for $i = 1, 2, \ldots, k$. This gives $[B_q]_{\mathcal{B}} = VDV^T$. It turns out that, under a certain condition on V, the matrices $[B_q]_{\mathcal{B}}$ and D have the same rank and signature.

Proposition 5.2.20 ([8, Proposition 8.13])**.** *Let S_i, $i = 1, 2$, be real and symmetric matrices of order n_i. If there is a full rank $n_2 \times n_1$-matrix P such that $S_2 = PS_1P^t$, then*

$$\begin{aligned} \operatorname{rank}(S_1) &= \operatorname{rank}(S_2)\,, \ \textit{and} \\ \operatorname{sign}(S_1) &= \operatorname{sign}(S_2) \end{aligned}$$

Thus, if we can show that V has full rank, we can determine the rank and signature of B_q by computing the rank and signature of D.

Let us start with finding the rank and signature of D. Since D is a diagonal matrix with $d_i q(\alpha_i)$ on the diagonal, we get $\operatorname{rank}(D) = |\{j \mid q(\alpha_j) \neq 0\}|$. Now we turn to finding the signature of D. Since we can re-index the α_i, we can, without loss of generality, assume that

$$\alpha_i \in \begin{cases} \mathbb{R}^n, & 1 \leq i \leq r \\ \mathbb{C}^n \backslash \mathbb{R}^n, & r+1 \leq i \leq k \end{cases}.$$

Let A be a diagonal matrix with

$$a_{ii} = \begin{cases} (d_i \, |q(\alpha_i)|)^{-1/2}, & 1 \leq i \leq r \\ 1, & r+1 \leq i \leq k \end{cases}$$

and let $\mathcal{B}'$ be the basis such that A is the change of basis matrix from $\mathcal{B}$ to $\mathcal{B}'$. This gives

$$\begin{aligned} [D]_{\mathcal{B}'} &= A^T [D]_{\mathcal{B}} A \\ &= \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix} \end{aligned}$$

where D_1 is a diagonal matrix in which every diagonal element is either $1, -1$ or 0 — more precisely, the number of 1:s is $|\{j \mid 1 \leq j \leq r \text{ and } q(\alpha_j) > 0\}|$, and the number of -1:s is $|\{j \mid 1 \leq j \leq r \text{ and } q(\alpha) < 0\}|$ — and D_2 is a diagonal matrix with diagonal elements $q(\alpha_{r+1}),\ q(\alpha_{r+2}),\ \ldots, q(\alpha_k)$. Thus,

$$\begin{aligned} \operatorname{sign}(D_1) &= |\{j \mid 1 \leq j \leq r \text{ and } q(\alpha_j) > 0\}| - |\{j \mid 1 \leq j \leq r \text{ and } q(\alpha) < 0\}| \\ &= |\{\alpha \in \mathbb{R}^n \mid q(\alpha) > 0\}| - |\{\alpha \in \mathbb{R}^n \mid q(\alpha) < 0\}|. \end{aligned}$$

Next, let us consider D_2. Assume that $\alpha \in V(I) \cap (\mathbb{C}^n \backslash \mathbb{R}^n)$. Then $p(\alpha) = 0$ for every $p \in I$. Let $\overline{\alpha}$ denote the component-wise complex conjugate of α. Since

$$\begin{aligned} \overline{z_1 + z_2} &= \overline{z_1} + \overline{z_2}, \text{ and} \\ \overline{z_1 z_2} &= \overline{z_1}\ \overline{z_2} \end{aligned}$$

this implies that $p(\overline{\alpha}) = \overline{p(\alpha)} = 0$ for all $p \in I$. Thus, $\overline{\alpha} \in V(I)$. By the same argument as above, we can, without loss of generality, assume that $\alpha_{r+2} = \overline{\alpha_{r+1}}$, $\alpha_{r+4} = \overline{\alpha_{r+3}}$, and so on. Consider two consecutive terms of the sum $\sum_{s=r+1}^{n} d_s q(\alpha_s) v_i(\alpha_s) v_j(\alpha_s)$:

$$\begin{aligned} & d_{r+1} q(\alpha_{r+1}) v_i(\alpha_{r+1}) v_j(\alpha_{r+1}) + d_{r+1} q(\overline{\alpha_{r+1}}) v_i(\overline{\alpha_{r+1}}) v_j(\overline{\alpha_{r+1}}) \\ =& 2 d_{r+1} \operatorname{Re}\left(q(\alpha_{r+1}) v_i(\alpha_{r+1}) v_j(\alpha_{r+1})\right) \\ =& 2 d_{r+1} \left[\operatorname{Re}\left(q(\alpha_{r+1})\right) \operatorname{Re}\left(v_i(\alpha_{r+1})\right) \operatorname{Re}\left(v_j(\alpha_{r+1})\right)\right. \\ & \quad - \operatorname{Re}\left(q(\alpha_{r+1})\right) \operatorname{Im}\left(v_i(\alpha_{r+1})\right) \operatorname{Im}\left(v_j(\alpha_{r+1})\right) \\ & \quad - \operatorname{Im}\left(q(\alpha_{r+1})\right) \operatorname{Re}\left(v_i(\alpha_{r+1})\right) \operatorname{Im}\left(v_j(\alpha_{r+1})\right) \\ & \quad \left. - \operatorname{Im}\left(q(\alpha_{r+1})\right) \operatorname{Re}\left(v_j(\alpha_{r+1})\right) \operatorname{Im}\left(v_i(\alpha_{r+1})\right)\right] \\ =& w_{ij,r+1}^T M_{r+1} w_{ij,r+1} \end{aligned}$$

where

$$(w_{ij,r+1})^T = \begin{pmatrix}\operatorname{Re}(v_i(\alpha_{r+1})) & \operatorname{Im}(v_i(\alpha_{r+1})) & \operatorname{Re}(v_j(\alpha_{r+1})) & \operatorname{Im}(v_j(\alpha_{r+1}))\end{pmatrix}$$

and

$$M_{r+1} = d_{r+1}\begin{pmatrix} 0 & 0 & a_{r+1} & -b_{r+1} \\ 0 & 0 & -b_{r+1} & -a_{r+1} \\ a_{r+1} & -b_{r+1} & 0 & 0 \\ -b_{r+1} & -a_{r+1} & 0 & 0 \end{pmatrix}$$

where

$$\begin{aligned} a_{r+1} &= d_{r+1}\operatorname{Re}(q(\alpha_{r+1})) \\ b_{r+1} &= d_{r+1}\operatorname{Im}(q(\alpha_{r+1})) \end{aligned}$$

The matrix M_{r+1} has eigenvalues $\pm\sqrt{a_{r+1}^2 + b_{r+1}^2}$, each with multiplicity two. Let

$$w^T = \begin{pmatrix} w_{ij,r+1}^T & w_{ij,r+3}^T & \cdots & w_{ij,k-1}\end{pmatrix}$$

and let M be the block diagonal matrix with $M_{r+(2j+1)}$ as the j:th block, for $j = 1, 2, \ldots, (k-r-2)/2$; note that the last index is indeed an integer, since $k-r$, which is equal to the number of elements in $V(I) \cap (\mathbb{C}^n \backslash \mathbb{R}^n)$, must be even. Then

$$\sum_{s=r+1}^{k} d_s q(\alpha_s) v_i(\alpha_s) v_j(\alpha_s) = w^T M w.$$

The eigenvalues of M are

$$\left\{\pm\sqrt{a_{r+(2j+1)}^2 + b_{r+(2j+1)}^2} \mid 1 \le j \le (k-r-2)/2)\right\}.$$

Thus, the number of positive eigenvalues of M is equal to the number of negative eigenvalues of M (counted with multiplicity). Since M is real and symmetric, it is the matrix of a symmetric bilinear form $\tilde{M}$ in some basis $\mathcal{B}'$ of $\mathbb{R}[x]/I$, i.e. $\left[\tilde{M}\right]_{\mathcal{B}'} = M$. By Proposition 5.1.8, there is a a basis B'' such that $\left[\tilde{M}\right]_{\mathcal{B}''}$ is a diagonal matrix whose elements belong to $\{-1, 1, 0\}$. Thus,

$$\sum_{s=r+1}^{k} d_s q(\alpha_s) v_i(\alpha_s) v_j(\alpha_s) = \sum_{i=1}^{k} u_i^2 - \sum_{i=k+1}^{m} u_i^2$$

for some u_i, where $2k = m$. By the uniqueness part of Proposition 5.1.8, this implies that $\operatorname{sign}(D_2) = 0$. Since the set of eigenvalues of D is the union of the set of eigenvalues of D_1 and the set of eigenvalues of D_2, we conclude that

$$\operatorname{sign}(D) = |\{\alpha \in \mathbb{R}^n \mid q(\alpha) > 0\}| - |\{\alpha \in \mathbb{R}^n \mid q(\alpha) < 0\}|.$$

Now we turn to showing that V has full rank. Either I is radical or it is not. We consider the two cases separately.

First, consider the case where I is radical. Assume that there are λ_i, $i = 1, 2, \ldots, k$, such that $\{\lambda_1, \lambda_2, \ldots, \lambda_k\} \neq \{0\}$ and

$$\sum_{i=1}^{m} \lambda_i \begin{pmatrix} v_i(\alpha_1) \\ v_i(\alpha_2) \\ \vdots \\ v_i(\alpha_k) \end{pmatrix} = 0.$$

Let $h(x) = \sum_{i=1}^{k} \lambda_i v_i(x)$. Then $h(\alpha_i) = 0$ for all i, so $h \in \sqrt{I} = I$, by Hilbert's Nullstellensatz. In other words, $\sum_{i=1}^{k} \lambda_i v_i(x) \in I$, where not all λ_i are zero; but this contradicts that $\{v_1, v_2, \dots, v_k\}$ is a basis. Hence, we have shown for radical I that the rows of V are linearly independent, so D and B_q have the same rank and signature, by Proposition 5.2.20.

If I is not radical, then we must choose a certain basis of $\mathbb{R}[x]/I$ to get a V which we can easily verify has full rank. The basis in question is

$$\left\{1, v, v^2, \dots, v^{k-1}, v_{k+1}, v_{k+2}, \dots, v_n\right\}$$

[2, Chapter 4.5], where v is a polynomial such that

(i) $v(\alpha_i) \neq v(\alpha_j)$ for all distinct $\alpha_i, \alpha_j \in V(I)$,

(ii) $\left\{1, v, v^2, \dots, v^{k-1}\right\}$ is linearly independent, and

(iii) $v_i,\ i = k{+}1, k{+}2, \dots, n$, are chosen so that $\left\{1, v, v^2, \dots, v^{k-1}\right\}$ is extended to a basis of $\mathbb{R}[x]/I$.

Given a v with properties (i) and (ii), we can always extend $\left\{1, v, v^2, \dots, v^{n-1}\right\}$ to a basis of $\mathbb{R}[x]/I$; this is a result of elementrary linear algera. It remains to show that we can find a v with properties (i) and (ii).

Lemma 5.2.21 (main part of [2, Lemma 4.89])**.** *There is a constant c such that $v_c(x) = x_1 + cx_2 + c^2x_3 + \dots + c^{n-1}x_n$ satisfies that $v_c(\alpha_i) \neq v_c(\alpha_j)$ for all distinct $\alpha_i, \alpha_j \in V(I)$*

Proof. This proof follows the proof in the cited book. Let α_i, α_j be distinct elements of $V(I)$. We will use the notation $\alpha_k = (\alpha_{k1}, \dots, \alpha_{kn})$. This gives $v_c(\alpha_k) = \alpha_{k1} + c\alpha_{k2} + c^2\alpha_{k3} + \dots c^{n-1}\alpha_{kn}$, so

$$\begin{aligned} v_c(\alpha_i) - v_c(\alpha_j) &= \sum_{k=1}^{n} c^{k-1} \left(\alpha_{ik} - \alpha_{jk}\right) \\ &= \tilde{v}_{ij}(c) \end{aligned}$$

where

$$\tilde{v}_{ij}(t) = \sum_{k=1}^{n} \left(\alpha_{ik} - \alpha_{jk}\right) t^{k-1}.$$

This is a univariate polynomial of degree $(n-1)$. This means that there are at most $(n-1)$ choices of c such that $v_c(\alpha_i) = v_c(\alpha_j)$. We want to find a c such that $v_c(\alpha_i) \neq v_c(\alpha_j)$ for every choice of distinct $i, j \in \{1, 2, \dots, k\}$. Given distinct i, j, let c_{ijk}, $k = 1, 2, \dots, n-1$, be the constants which satisfies $c_{ijk}(\alpha_i) = c_{ijk}(\alpha_j)$. Then any choice of $c \notin \cup_{i,j} \{c_{ijk} \mid k \in \{1, 2, \dots, n-1\}\}$ satifies that $v_c(\alpha_i) \neq v_c(\alpha_j)$ for every choice of distinct i, j. □

Lemma 5.2.22 ([2, Lemma 4.90])**.** *Any polynomial which satisfies (i) will also satisfy (ii).*

Proof. This proof follows the proof in the cited book. Let v be a polynomial which satifies (i). Assume that there exists $\lambda_i,\ i = 0, 1, \dots, k-1$, such that $\{\lambda_1, \lambda_2, \dots, \lambda_{k-1}\} \neq \{0\}$ and $\sum_{i=0}^{k-1} \lambda_i v^i \in I$. Then

$$\sum_{j=0}^{k-1} \lambda_j v(\alpha_i)^j = 0$$

for all $\alpha_i \in V(I)$, i.e. for $i = 1, 2, \ldots, k$. But $\sum_{j=0}^{k-1} \lambda_j t^j$ is a univariate polynomial of degree $(k-1)$, so it has at most $k-1$ distinct roots, which means we have a contradiction. Hence, $\sum_{j=0}^{k-1} \lambda_j v^j \equiv 0$, so $\left\{1, v, v^2, \ldots, v^{k-1}\right\}$ is a linearly independent set. □

Let $\mathcal{B} = \left\{1, v, v^2, \ldots, v^{k-1}, v_{k+1}, v_{k+2}, \ldots, v_n\right\}$ be a basis of $\mathbb{R}[x]/I$ satisfying (i)-(iii). Then

$$V = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ v(\alpha_1) & v(\alpha_2) & \ldots & v(\alpha_k) \\ v(\alpha_1)^2 & v(\alpha_2)^2 & \ldots & v(\alpha_k)^2 \\ v(\alpha_1)^3 & v(\alpha_2)^3 & \ldots & v(\alpha_k)^3 \\ \vdots & \vdots & \ddots & \vdots \\ v(\alpha_1)^{k-1} & v(\alpha_2)^{k-1} & \ldots & v(\alpha_k)^{k-1} \\ v_{k+1}(\alpha_1) & v_{k+1}(\alpha_2) & \ldots & v_{k+1}(\alpha_k) \\ \vdots & \vdots & \ddots & \vdots \\ v_n(\alpha_1) & v_n(\alpha_2) & \ldots & v_n(\alpha_k) \end{pmatrix}.$$

The matrix $\tilde{V}$ consisting of the first k rows of V is a so called *Vandermonde matrix*; the determinant of such a matrix has a certain form.

Proposition 5.2.23 ([8, Proposition 3.19]). *For any choice of elements $c_i \in R$, $i = 1, 2, \ldots, n$, we have* $\det\left(\left(c_j^{i-1}\right)_{i,j}\right) = \prod_{1 \leq k < m \leq n} (c_k - c_m)$.

Since $\alpha_i \neq \alpha_j$ for $i \neq j$, we have $\det \tilde{V} \neq 0$, by Proposition 5.2.23. Since V is an $n \times k$-matrix, this implies that it has full rank.

6 Determining the stability properties of steady states

Let α be a steady state of a dynamical system $\dot{x}_i = f_i(x),\ i = 1, 2, \ldots, n$. It is of great interest to determine how the system will behave when it is in a state close to, but not equal to, α. Assume that the system is in a state close to α. Will the future states of the system still be close to α? Then we say that α is *stable*. Will the future states not only be close to, but even closer and closer to, α? Then we say that α is *asymptotically stable*. If α is not stable, we say that it is *unstable*. If α is unstable, there exists states close to it, such that if the system is in such a state at one point in time, it will still stray far away from α in the future. Let us record what has been said in a precise definition.

Definition 6.0.1 ([13, Definition 1 in chapter 2.9]). *Let $\dot{x}_i = f_i(x),\ i = 1, 2, \ldots, n$, be a dynamical system. Let $x(t)$ be a function satisfying these dynamics. Let $\alpha \in \mathbb{R}^n$ be a steady state of the system. Let*

$$N_\delta(\alpha) = \{x \in \mathbb{R}^n \mid |x - \alpha| < \delta\}.$$

If

$$\forall \epsilon > 0\ \exists \delta > 0 :\ x(0) \in N_\delta(\alpha) \ \Rightarrow\ x(t) \in N_\epsilon(\alpha)$$

we say that α is a stable steady state. If α is not stable, we say that it is unstable.

Assume that there is a $\delta > 0$ such that the following implication holds:

$$x(0) \in N_\delta(\alpha) \;\Rightarrow\; \lim_{t\to+\infty} x(t) = \alpha.$$

Then we say that α is an asymptotically stable steady state.

Let us make a remark on the importance of stability of a steady state for mathematical modelling. Consider a dynamical system which is supposed to model some phenomenon. For this argument, let us call the phenomenon the "true system" and the dynamical system which is supposed to model it the "model system". A steady state of the model system corresponds to a steady state of the true system with the same stability properties. Assume that the current state of the system is a steady state. If the model system is a perfect model of the true system, then we know that the true system will stay in the steady state forver. But, of course, any model will be only an approximation of the phenomenon, so the steady state $\hat{\alpha}$ of the true system is not exactly the same as the steady state α of the model system. If α, and therefore $\hat{\alpha}$, is unstable, this implies that the future states of the true system can lie very far from α, even though, according to the model, the system should stay in α. On the other hand, if α is stable, it might still be that the true system is in a non-steady state, but it will at least be close to $\hat{\alpha}$. This implies that it will stay close to $\hat{\alpha}$, since α, and therefore $\hat{\alpha}$, is assumed to be stable.

There is a couple of very useful results for determining the stability properties of a steady state. Before we can formulate them, we must recall the concept of the *Jacobian* of a function.

Definition 6.0.2. *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be differentiable. Then the function $J_f : \mathbb{R}^n \to \mathbb{R}^{nm}$ defined by*

$$J_f(x) = \left(\frac{\partial f_i}{\partial x_j}(x)\right)_{\substack{1\leq i\leq m\\ 1\leq j\leq n}},$$

is called the Jacobian of f.

Now we can formulate the results.

Proposition 6.0.3 ([10, The Theorem in §2 of chapter 9])**.** *Let α be a steady state of a dynamical system $\dot{x}_i = f_i(x)$, $i = 1, 2, \ldots, n$, where $f : \mathbb{R}^n \to \mathbb{R}^n$ is continuously differentiable. If α is stable, then $\mathrm{Re}\,(\lambda_i) \leq 0$ for all eigenvalues λ_i of $J_f(\alpha)$.*

Proposition 6.0.4 ([18, follows from Theorem 6.10])**.** *Let α be a steady state of a dynamical system $\dot{x}_i = f_i(x)$, $i = 1, 2, \ldots, n$, where $f : \mathbb{R}^n \to \mathbb{R}^n$ is continuously differentiable. If $\mathrm{Re}\,(lambda_i) < 0$ for all eigenvalues λ_i of $J_f(\alpha)$, then α is asymptotically stable.*

Corollary 6.0.5. *Let α be a steady state of a dynamical system $\dot{x}_i = f_i(x)$, $i = 1, 2, \ldots, n$, where $f : \mathbb{R}^n \to \mathbb{R}^n$ is continuously differentiable, such that $\mathrm{Re}\,(\lambda_i) \neq 0$ for all eigenvalues λ_i of $J_f(\alpha)$. Then α is asymptotically stable if and only if $\mathrm{Re}\,(\lambda_i) < 0$ for all i.*

Proof. Assume that α is asymptotically stable. Then α is stable. By Proposition 6.0.3, this means that $\mathrm{Re}\,(\lambda_i) \leq 0$ for all eigenvalues λ_i of $J_f(\alpha)$. Since $\mathrm{Re}\,(\lambda_i) \neq 0$ by assumption, this implies that $\mathrm{Re}\,(\lambda_i) < 0$ for all eigenvalues λ_i of $J_f(\alpha)$.

Conversely, assume that $\mathrm{Re}\,(\lambda_i) < 0$ for all eigenvalues λ_i of $J_f(\alpha)$. Then Proposition 6.0.4 implies that α is asymptotically stable. □

Let $\dot{x}_i = p_i(x),\ i = 1, 2, \ldots, n$, be a polynomial dynamical system. Fix a monomial order $<$ and let

$$\dot{x}_i = C_< m_<$$

be the matrix representation of this system corresponding to $<$. The right-hand side depends on $x = (x_1, x_2, \ldots, x_n)$; more precisely, $m_<$ but not $C_<$ depends on x. This gives

$$J_{C_< m_<}(x) = C_< J_{m_<}(x). \tag{6.1}$$

Let α be a steady state of the system. Proposition 6.0.3 and Corollary 6.0.5 suggests that, to determine the stability properties of α, it is useful to know something about the distribution of eigenvalues of $C_< J_{m_<}(\alpha)$ in the complex plane. The following result can be used for this purpose.

Proposition 6.0.6 ([9, Theorem 2 in chapter XV])**.** *Let $f \in \mathbb{R}[t]$. Write $f(t) = a_{10}t^n + a_{20}t^{n-1} + a_{11}t^{n-2} + a_{21}t^{n-3} + \ldots$. Let*

$$\begin{aligned} f_1(t) &= a_{10}t^n - a_{11}t^{n-2} + a_{12}t^{n-4} - \ldots \\ f_2(t) &= a_{20}t^{n-1} - a_{21}t^{n-3} + a_{22}t^{n-5} - \ldots \end{aligned}$$

and, for $i \geq 3$, define $(-f_i)$ as the remainder when f_{i-2} is divided by f_{i-1}. Let k be the number such that $f_k \neq 0$ but $f_{k+1} = 0$. Then

(i) *$f_i = a_{i0}t^{n-(i-1)} + a_{i1}t^{n-(i-1)-2} + \ldots$ for some a_{ij},*

(ii) *the number of sign changes in the sequence $(a_{10}, a_{20}, a_{30}, \ldots, a_{k0})$ is equal to*

$$|\{t \in \mathbb{C} \mid f(t) = 0 \text{ and } \mathrm{Re}\,(t) > 0\}|,$$

and

(iii) *all roots x of f satisfy $\mathrm{Re}\,(x) < 0$ if and only if $(a_{10}, a_{20}, a_{30}, \ldots, a_{k0})$ has no zeros and no sign changes.*

Example 6.0.7. We once again return to Example 3.8.6. Let $<$ be the monomial ordering with $x_1 > x_2 > x_3$. Then the system has the matrix representation $\dot{x}_i = C_< m_<$, with

$$C_< = \begin{pmatrix} 1 & -1 & -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 2 & 1 & -4 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \end{pmatrix}$$

and

$$m_<^T = \begin{pmatrix} x_1^2x_2^3x_3 & x_1x_2^2x_3 & x_1x_2x_3^2 & x_2^2x_3^4 & x_2x_3^3 & x_2x_3^2 & x_3^6 & 1 \end{pmatrix}.$$

We compute

$$J_{m_<}(x_1, x_2, x_3) = \begin{pmatrix} 2x_1x_2^3x_3 & 3x_1^2x_2^2x_3 & x_1^2x_2^3 \\ x_2^2x_3 & 2x_1x_2x_3 & x_1x_2^2 \\ x_2x_3^2 & x_1x_3^2 & 2x_1x_2x_3 \\ 0 & 2x_2x_3^4 & 4x_2^2x_3^3 \\ 0 & x_3^3 & 3x_2x_3^2 \\ 0 & x_3^2 & 2x_2x_3 \\ 0 & 0 & 6x_3^5 \\ 0 & 0 & 0 \end{pmatrix}.$$

Recall that the steady states of the system are $\alpha_1 = (1, 1, 1)$ and $\alpha_2 = (1, -1, 1)$. The characteristic polynomial of $C_< J_{m_<}(\alpha_1)$ is

$$-t^3 + 6t^2 + 6t.$$

We see immediately that 0 is an eigenvalue of $J_{m_<}(\alpha_1)$. Therefore, we will not be able to use Corollary 6.0.5; we can use Proposition 6.0.3, however. Let

$$\begin{aligned} f_1(t) &= -t^3 - 6t, \text{ and} \\ f_2(t) &= 6t^2. \end{aligned}$$

We define $f_3(t)$ as (-1) times the remainder of $f_1(t)$ divided by $f_2(t)$. Since

$$-t^3 - 6t = 6t^2 \cdot \left(-\frac{1}{6}t\right) - 6t$$

we have $f_3(t) = 6t$. Since

$$6t^2 = 6t \cdot t$$

we have that $f_2(t)$ divided by $f_3(t)$ leaves no remainder. Therefore, we shall consider the sequence $(-1, 6, 6)$. Since it has one sign change, $C_< J_{m_<}(\alpha_1)$ has two eigenvalues with positive real part. Hence, α_1 is an unstable steady state.

In this case, we can easily solve for the eigenvalues explicitly; let us do this and compare with what was just said. The roots of $-t^3 + 6t^2 + 6t$ are $t = 0$ and $t = 3 \pm \sqrt{15}$. Hence, there is exactly one eigenvalue with positive real part, which is what we expected.

The characteristic polynomial of $C_< J_{m_<}(\alpha_2)$ is

$$-t^3 - 4t^2 + 2t + 12.$$

We define

$$\begin{aligned} f_1(t) &= -t^3 - 2t, \text{ and} \\ f_2(t) &= -4t^2 - 12. \end{aligned}$$

This gives $f_3(t) = -t$ and $f_4(t) = -12$. We shall therefore consider the sequence $(-1, -4, -1, 12)$. This sequence has one sign change: hence, there is one eigenvalue with positive real part. Hence, the point $(1, -1, 1)$ is an unstable steady state.

Again, let us find the roots the characteristic polynomial. We can easily find the root $t = -2$, and can then easily find the rest of the roots. This gives the roots $t = -2$ and $t = -1 \pm \sqrt{7}$. Hence, there is exactly one eigenvalue with positive real part, which is what we expected. ◇

7 The class of chemical reaction networks

So far, we have considered general polynomial dynamical systems. Now we will consider some particular classes of polynomial dynamical systems. In this section, we study so called *chemical reaction networks.*

7.1 Introduction

A chemical reaction network is a set of chemical substances together with a set of possible reactions among the substances. Each reaction in a chemical reaction network involving the substrates $x_1, x_2, \dots, x_n$ can schematically be written

$$\textstyle\sum_{j=1}^n c_{ji} x_j \xrightarrow{\gamma_i} \sum_{j=1}^n d_{ji} x_j,$$

where the $c_{ji}, d_{ji} \in \mathbb{R}$ are non-negative and the γ_i are positive. The left-hand side $\sum_{j=1} c_{ji} x_j$ and right-hand side $\sum_{j=1}^n d_{ji} x_j$ are called *complices* (plural of *complex*). Let r be the number of reactions in the network. Under the assumption of law of mass action (see Example 2.1.1), the dynamics of the concentrations of $x_1, x_2, \dots, x_n$ are then described by

$$\dot{x_j} = \sum_{i=1}^{r} (d_{ji} - c_{ji}) \gamma_i \prod_{k=1}^{n} x_k^{c_{ki}},\ j = 1, 2, \dots, n \tag{7.1}$$

which is a polynomial dynamical system. Let $p_j = \sum_{i=1}^r (d_{ji} - c_{ji}) \gamma_i \prod_{k=1}^n x_k^{c_{ki}}$ and let $P = \{p_1, p_2, \dots, p_n\}$. Let $<$ be a monomial ordering and μ an enumeration of P. We can reindex so that $\gamma_i \prod_{k=1} x_k^{c_{ki}} < \gamma_{i+1} \prod_{k=1} x_k^{c_{k,i+1}}$ for all i. Then

$$C_{<,\mu} = (d_{ij} - c_{ij})_{\substack{1 \le i \le n \\ 1 \le j \le r}}$$

is the matrix representation of this system. In the context of chemical reaction networks, this matrix is called a stoichiometric matrix.

7.2 Application of the theory of polynomial dynamical systems

Example 7.2.1. Consider the chemical reaction network

$$S_0 + E \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES_0 \xrightarrow{k_2} S_1 + E \underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} ES_1 \xrightarrow{k_4} S_2 + E$$

$$S_2 + F \underset{k_{-5}}{\overset{k_5}{\rightleftharpoons}} FS_2 \xrightarrow{k_6} S_1 + F \underset{k_{-7}}{\overset{k_7}{\rightleftharpoons}} FS_1 \xrightarrow{k_8} S_0 + F.$$

Let $C_i = ES_i$ for $i = 0, 1$, and $D_i = FS_i$ for $i = 1, 2$. Under the assumption of law of mass action, the dynamics of this network is given by

$$\begin{cases} \dot{C_0} = & k_1 S_0 E - (k_{-1} + k_2) C_0 \\ \dot{C_1} = & k_3 S_1 E - (k_{-3} + k_4) C_1 \\ \dot{D_1} = & k_7 S_1 F - (k_{-7} + k_8) D_1 \\ \dot{D_2} = & k_5 S_2 F - (k_{-5} + k_6) D_2 \\ \dot{S_0} = & -k_1 S_0 E + k_{-1} C_0 + k_8 D_1 \\ \dot{S_1} = & -k_3 S_1 E - k_7 S_1 F + k_2 C_0 + k_{-3} C_1 + k_{-7} D_1 + k_6 D_2 \\ \dot{S_2} = & -k_5 S_2 F + k_4 C_1 + k_{-5} D_2 \\ \dot{E} = & -k_1 S_0 E - k_3 S_1 E + (k_{-1} + k_2) C_0 + (k_{-3} + k_4) C_1 \\ \dot{F} = & -k_7 S_1 F - k_5 S_2 F + (k_{-7} + k_8) D_1 + (k_{-5} + k_6) D_2 \end{cases} . \tag{7.2}$$

Note that the right-hand sides are elements in $\mathbb{R}[E, F, S_0, S_1, S_2, C_0, C_1, D_1, D_2]$. Let I be the ideal generated by the polynomials in the right-hand side. Order the monomials in this polynomial ring according to Lex with

$$E > F > S_0 > S_1 > S_2 > C_0 > C_1 > D_1 > D_2.$$

Then

$$m_<^T = \begin{pmatrix} ES_0 & ES_1 & FS_1 & FS_2 & C_0 & C_1 & D_1 & D_2 \end{pmatrix}.$$

The coefficient matrix corresponding to the ordering $<$ is $C_< =$

$$\begin{pmatrix}
k_1 & 0 & 0 & 0 & -(k_{-1}+k_2) & 0 & 0 & 0 \\
0 & k_3 & 0 & 0 & 0 & -(k_{-3}+k_4) & 0 & 0 \\
0 & 0 & k_7 & 0 & 0 & 0 & -(k_{-7}+k_8) & 0 \\
0 & 0 & 0 & k_5 & 0 & 0 & 0 & -(k_{-5}+k_6) \\
-k_1 & 0 & 0 & 0 & k_{-1} & 0 & k_8 & 0 \\
0 & -k_3 & -k_7 & 0 & k_2 & k_{-3} & k_{-7} & k_6 \\
0 & 0 & 0 & -k_5 & 0 & k_4 & 0 & k_{-5} \\
-k_1 & -k_3 & 0 & 0 & k_{-1}+k_2 & k_{-3}+k_4 & 0 & 0 \\
0 & 0 & -k_7 & -k_5 & 0 & 0 & k_{-7}+k_8 & k_{-5}+k_6
\end{pmatrix}. \tag{7.3}$$

First, let us investigate whether the dimension of the system can be reduced. We compute the rank of $C_<$; the rank is 6. By Proposition 2.3.5, this implies that $\dim\ker\left(C_{<,\mu}\right)^T = 9 - 6 = 3$. Thus, there are three linearly independent conservation laws of this system. Hence, (7.2) can be reduced to a system of dimension 6. A basis for the kernel is

$$\left\{(0,0,1,1,0,0,0,0,1)^T,\ (1,1,0,0,0,0,0,1,0)^T,\ (1,1,1,1,1,1,1,0,0)^T\right\}.$$

This gives three conservation laws,

$$\begin{cases}
\phantom{C_0 + C_1 +{}} D_1 + D_2 \phantom{{}+ S_0 + S_1 + S_2} + F = a_1 \\
C_0 + C_1 \phantom{{}+ D_1 + D_2 + S_0 + S_1 + S_2} + E = a_2 \\
C_0 + C_1 + D_1 + D_2 + S_0 + S_1 + S_2 = a_3
\end{cases}$$

where $a_1, a_2, a_3 \in \mathbb{R}$. By solving for three of the variables, and substituting those variables in (7.2), we get a reduced system. For example, by solving for (F, E, S_2) and substituting into (7.2), we get

$$\left\{\begin{array}{rcl}
\dot{C}_0 & = & -k_1S_0C_0 - k_1S_0C_1 + k_1a_2S_0 - (k_{-1}+k_2)C_0 \\
\dot{C}_1 & = & -k_3S_1C_0 - k_3S_1C_1 + k_3a_2S_1 - (k_{-3}+k_4)C_1 \\
\dot{D}_1 & = & -k_7S_1D_1 - k_7S_1D_2 + k_7a_1S_1 - (k_{-7}+k_8)D_1 \\
\dot{D}_2 & = & -k_5S_0D_1 - k_5S_0D_2 + k_5a_1S_0 - k_5S_1D_1 - k_5S_1D_2 + k_5a_1S_1 \\
 & & -k_5C_0D_1 - k_5C_0D_2 + k_5a_1C_0 - k_5C_1D_1 - k_5C_1D_2 + k_5a_1C_1 \\
 & & -k_5D_1^2 - 2k_5D_1D_2 + k_5(a_3+a_1)D_1 - k_5D_2^2 \\
 & & +(k_5a_3 - (k_{-5}+k_6))D_2 - k_5a_1a_3 \\
\dot{S}_0 & = & k_1S_0C_0 + k_1S_0C_1 - k_1a_2S_0 + k_{-1}C_0 + k_8D_1 \\
\dot{S}_1 & = & k_3S_1C_0 + k_3S_1C_1 + k_7S_1D_1 + k_7S_1D_2 - (k_7a_1 + k_3a_2)S_1 \\
 & & +k_2C_0 - k_{-3}C_1 + k_{-7}D_1 + k_6D_2
\end{array}\right. . \tag{7.4}$$

The right-hand sides are polynomials in $\mathbb{R}[S_0, S_1, C_0, C_1, D_1, D_2]$.

Let us choose the parameters

$$k_i = \begin{cases} 2, & i \in \{1,3,5,7\} \\ 1, & \text{otherwise} \end{cases}, \text{ and}$$
$$k_{-i} = 1, \text{ for } i = 1,3,5,7.$$

Let $S_0(0) = E(0) = F(0) = 1$ and $S_1(0) = D_1(0) = D_2(0) = 1$; this gives $a_1 = a_2 = a_3 = 1$. Then

$$\begin{cases} \dot{C}_0 = & -2S_0C_0 - 2S_0C_1 + 2S_0 - 2C_0 \\ \dot{C}_1 = & -2S_1C_0 - 2S_1C_1 + 2S_1 - 2C_1 \\ \dot{D}_1 = & -2S_1D_1 - 2S_1D_2 + 2S_1 - 2D_1 \\ \dot{D}_2 = & -2S_0D_1 - 2S_0D_2 + 2S_0 - 2S_1D_1 - 2S_1D_2 + 2S_1 \\ & -2C_0D_1 - 2C_0D_2 + 2\alpha_1C_0 - 2C_1D_1 - 2C_1D_2 + 2C_1 \\ & -2D_1^2 - 4D_1D_2 + 4D_1 - 2D_2^2 - 2 \\ \dot{S}_0 = & 2S_0C_0 + 2S_0C_1 - 2S_0 + C_0 + D_1 \\ \dot{S}_1 = & 2S_1C_0 + 2S_1C_1 + 2S_1D_1 + 2S_1D_2 - 4S_1 \\ & +C_0 - C_1 + D_1 + D_2 \end{cases}.$$

The set $\mathcal{G} = \{g_1, g_2, g_3, g_3, g_4, g_5, g_6\}$, with

$$\begin{aligned} g_1 &= 8D_2^5 + 184D_2^4 + 60D_2^3 - 342D_2^2 + 189D_2 - 81, \\ g_2 &= 2709D_1 - 16D_2^4 - 444D_2^3 - 1928D_2^2 + 255D_2 - 747, \\ g_3 &= 3C_1 - D_2, \\ g_4 &= 2709C_0 - 16D_2^4 - 444D_2^3 - 1928D_2^2 + 255D_2 - 747, \\ g_5 &= 16254S_1 + 664D_2^4 + 15416D_2^3 + 8976D_2^2 - 11034D_2 + 5265, \text{ and} \\ g_6 &= 1161S_0 + 64D_2^4 + 1604D_2^3 + 3498D_2^2 - 2052D_2 + 1053, \end{aligned}$$

is a Gröbner-basis of I with respect to $S_0 > S_1 > C_0 > C_1 > D_1 > D_2$. Note that

$$\begin{aligned} \mathcal{G} \cap \mathbb{R} &= \emptyset, \\ \mathcal{G} \cap \mathbb{R}[D_2] &= \{g_1\}, \\ \mathcal{G} \cap \mathbb{R}[D_1, D_2] \setminus \mathbb{R}[D_2] &= \{g_2\}, \\ \mathcal{G} \cap \mathbb{R}[C_1, D_1, D_2] \setminus \mathbb{R}[D_1, D_2] &= \{g_3\}, \\ \mathcal{G} \cap \mathbb{R}[C_0, C_1, D_1, D_2] \setminus \mathbb{R}[C_1, D_1, D_2] &= \{g_4\}, \\ \mathcal{G} \cap \mathbb{R}[S_1, C_0, C_1, D_1, D_2] \setminus \mathbb{R}[C_0, C_1, D_1, D_2] &= \{g_5\}, \text{ and} \\ \mathcal{G} \cap \mathbb{R}[S_0, S_1, C_0, C_1, D_1, D_2] \setminus \mathbb{R}[S_1, C_0, C_1, D_1, D_2] &= \{g_6\}, \end{aligned}$$

so $\mathcal{G}$ admits a strongly triangular form. Thus, we can use the algorithm of Corollary 3.8.5.

First, we shall solve $g_1(D_2) = 0$. We do this numerically, and get the three real solutions

$$\begin{aligned} D_2 &\approx -22.5820, \\ D_2 &\approx -1.84959, \text{ and} \\ D_2 &\approx 0.957006. \end{aligned}$$

In this particular case, it turns out that g_i for $i = 2, \dots, 6$ each depend on D_2 and just one other variable — and the last dependence is even linear. In general, we would have to solve $g_2(D_1, \alpha) = 0$ for D_1, for each solution α of $g_1(D_2) = 0$, then solve $g_3(C_1, \beta, \alpha) = 0$ for C_1, for each solution α of $g_1(D_2) = 0$ and each solution β of $g_2(D_1, \alpha)$, and so on. In each step, we might be required to solve a non-linear polynomial (univariate) equation. In our case, after substituting D_2 in the equations $g_i = 0, i = 2, 3, \dots, 6$, we get a linear system of equations.

- $D_2 \approx -22.5820$: This gives

$$\begin{cases} g_2 &= & -37472.58 &+& 2709D_10 \\ g_3 &= & 22.58195324 &+& 3C_1 \\ g_4 &= & -37472.58 &+& 2709C_0 \\ g_5 &= & -23062 &+& 16254S_1 \\ g_6 &= & 3027.1 &+& 1161S_0 \end{cases}.$$

  The system of equations $g_i = 0$, for $i = 2, \dots, 6$ has the solution

  $(S_0, S_1, C_0, C_1, D_1) = (2.6073,\ 1.4188,\ 13.83262,\ -7.52731775,\ 13.83262)$

  so

$$\alpha_1 \approx (2.607,\ 1.419,\ 13.83,\ -7.527,\ 13.83,\ -22.58)$$

  is an element in $V_{\mathbb{R}}(I)$.

- $D_2 \approx -1.84949$: This gives

$$\begin{cases} g_2 &= & -5192.15751 &+& 2709D_1 \\ g_3 &= & 1.849583606 &+& 3C_1 \\ g_4 &= & -5192.15751 &+& 2709C_0 \\ g_5 &= & -33391.8312 &+& 16254S_1 \\ g_6 &= & 7414.78122 &+& 1161S_0 \end{cases}.$$

  This gives that

$$\alpha_2 \approx (-6.387,\ 2.054,\ 1.917,\ -0.6165,\ 1.917,\ -1.849)$$

  is an element in $V_{\mathbb{R}}(I)$.

- $D_2 = 0.957006$: This gives

$$\begin{cases} g_2 &= & -2671.321405 &+& 2709D_1 \\ g_3 &= & -0.9570058362 &+& 3C_1 \\ g_4 &= & -2671.321405 &+& 2709C_0 \\ g_5 &= & 16994.99167 &+& 16254S_1 \\ g_6 &= & 3752.46567 &+& 1161S_0 \end{cases}.$$

  This gives that

$$\alpha_3 \approx (-3.232,\ -1.046,\ 0.9861,\ 0.3190,\ 0.9861,\ 0.9570)$$

  is an element in $V_{\mathbb{R}}(I)$.

Thus,

$$V_{\mathbb{R}}\left(\langle\mathcal{G}\rangle\right)=\{\alpha_1,\alpha_2,\alpha_3\}$$

where

$$\begin{array}{lllllllll} \alpha_1 & \approx & (-2.607, & 1.419, & 13.83, & -7.527, & 13.83, & -22.58) \\ \alpha_2 & \approx & (-6.387, & 2.054, & 1.917, & -0.6165, & 1.917, & -1.849) & . \\ \alpha_3 & \approx & (-3.232, & -1.046, & 0.9861, & 0.3190, & 0.9861, & 0.9570) \end{array}$$

Note, however, that neither of these points makes sense in the chemical reaction network settings, since all concentrations must be non-negative. Thus, the chemical reaction network given by (7.2) does not have any steady states.

By the Gröbner basis computed above, we have

$$\begin{aligned} \ell\left(\langle G\rangle\right) &= \langle\{\operatorname{lm}(g) \mid g\in G\}\rangle \\ &= \langle D_2^5, D_1, C_1, C_0, S_1, S_0\rangle, \end{aligned}$$

so

$$\mathcal{B}=\left\{1+I, D_2+I, D_2^2+I, D_2^3+I, D_2^4+I\right\}$$

is a basis of $\mathbb{R}[x]/I$, by Proposition 3.8.1.

Let Q_1 be the quadratic form induced by the constant polynomial 1. The matrix for Q_1 in the basis $\mathcal{B}$ is given by

$$\begin{pmatrix} \operatorname{tr}(T_1) & \operatorname{tr}(T_{D_2}) & \operatorname{tr}\left(T_{D_2^2}\right) & \operatorname{tr}\left(T_{D_2^3}\right) & \operatorname{tr}\left(T_{D_2^4}\right) \\ \operatorname{tr}(T_{D_2}) & \operatorname{tr}\left(T_{D_2^2}\right) & \operatorname{tr}\left(T_{D_2^3}\right) & \operatorname{tr}\left(T_{D_2^4}\right) & \operatorname{tr}\left(T_{D_2^5}\right) \\ \operatorname{tr}\left(T_{D_2^2}\right) & \operatorname{tr}\left(T_{D_2^3}\right) & \operatorname{tr}\left(T_{D_2^4}\right) & \operatorname{tr}\left(T_{D_2^5}\right) & \operatorname{tr}\left(T_{D_2^6}\right) \\ \operatorname{tr}\left(T_{D_2^3}\right) & \operatorname{tr}\left(T_{D_2^4}\right) & \operatorname{tr}\left(T_{D_2^5}\right) & \operatorname{tr}\left(T_{D_2^6}\right) & \operatorname{tr}\left(T_{D_2^7}\right) \\ \operatorname{tr}\left(T_{D_2^4}\right) & \operatorname{tr}\left(T_{D_2^5}\right) & \operatorname{tr}\left(T_{D_2^6}\right) & \operatorname{tr}\left(T_{D_2^7}\right) & \operatorname{tr}\left(T_{D_2^8}\right) \end{pmatrix}$$

where T_{D_i} is the linear transformation induced by multiplication with D_i.

First, since $(1+I)\cdot(f+I)=f+I$, we have $T_1=I$. This implies that $\operatorname{tr}(T_1)=5$, since $\mathbb{R}[S_0,S_1,C_0,C_1,D_1,D_2]/I$ has dimension five.

Next, take $f+I\in\mathbb{R}[x]/I$; then $f+I=\sum_{i=0}^{4}a_iD_2^i$ for some $a_i\in\mathbb{R}$. Multiplying with D_2+I gives

$$(D_2+I)(f+I)=a_0D_2+a_1D_2^2+a_2D_2^3+a_3D_2^4+a_4D_2^5.$$

Since

$$8D_2^5+184D_2^4+60D_2^3-342D_2^2+189D_2-81\in I,$$

we have the identity

$$D_2^5=-\frac{184}{8}D_2^4-\frac{60}{8}D_2^3+\frac{342}{8}D_2^2-\frac{189}{8}D_2+\frac{81}{8}$$

in $\mathbb{R}[S_0,S_1,C_0,C_1,D_1,D_2]/I$. This gives

$$\begin{aligned} (D_2+I)(f+I)=\frac{81}{8}a_4+\left(a_0-\frac{189}{8}a_4\right)D_2+\left(a_1+\frac{342}{8}a_4\right)D_2^2 \\ +\left(a_2-\frac{60}{8}\right)D_2^3+\left(a_3-\frac{184}{8}\right)D_2^4. \end{aligned}$$

Thus, in the basis $\mathcal{B}$, the linear transformation T_{D_2} has the matrix

$$C = \begin{pmatrix} 0 & 0 & 0 & 0 & 81/8 \\ 1 & 0 & 0 & 0 & -189/8 \\ 0 & 1 & 0 & 0 & 342/8 \\ 0 & 0 & 1 & 0 & -60/8 \\ 0 & 0 & 0 & 1 & -184/8 \end{pmatrix},$$

so $\operatorname{tr}(T_{D_2}) = -\frac{184}{8}$.

Note that $T_{D_2^2}(f) = D_2 T_{D_2}(f)$ for every f. There are $b_i \in \mathbb{R}$ such that $T_{D_2}(f) = b_0 + b_1 D_2 + b_2 D_2^2 + b_3 D_2^3 + b_4 D_2^4$. Let

$$a = \begin{pmatrix} a_0 & a_1 & a_2 & a_3 & a_4 \end{pmatrix}^T$$

and

$$b = \begin{pmatrix} b_0 & b_1 & b_2 & b_3 & b_4 \end{pmatrix}^T.$$

Then

$$\left[T_{D_2^2}(f)\right]_{\mathcal{B}} = Cb.$$

Since

$$[T_{D_2}(f)]_{\mathcal{B}} = Ca,$$

we have

$$\left[T_{D_2^2}(f)\right]_{\mathcal{B}} = C^2 a.$$

This means that $T_{D_2^2}$ has the matrix C^2 in basis $\mathcal{B}$. More generally, $T_{D_2^j}$ has the matrix C_j in basis $\mathcal{B}$.

We compute the powers of C and take the trace; we get

$$\begin{aligned} \operatorname{tr}\left(T_{D_2^2}\right) &= 514, \\ \operatorname{tr}\left(T_{D_2^3}\right) &= -46085/4, \\ \operatorname{tr}\left(T_{D_2^4}\right) &= 260056, \\ \operatorname{tr}\left(T_{D_2^5}\right) &= -46978489/8, \\ \operatorname{tr}\left(T_{D_2^6}\right) &= 2121725221/16, \\ \operatorname{tr}\left(T_{D_2^7}\right) &= -95825372287/32, \text{ and} \\ \operatorname{tr}\left(T_{D_2^8}\right) &= 1081962013723/16. \end{aligned}$$

Thus, we have computed Q_1. The characteristic polynomial of Q_1 is

$$\begin{aligned} \chi_{Q_1}(\lambda) \approx -\lambda^5 &+ 6.77555 \cdot 10^{10}\lambda^4 - 3.72156 \cdot 10^{12}\lambda^3 \\ &+ 1.73569 \cdot 10^{13}\lambda^2 - 1.0956 \cdot 10^{13}\lambda - 4.24975 \cdot 10^{12}. \end{aligned}$$

It follows from Descartes' rule of sign that the number of positive eigenvalues is 4, while the number of negative eigenvalues is 1. Thus, the signature is 3, so this system has three steady states.

Let us compute the eigenvalues, to check that what was just said is correct. The eigenvalues of Q_1 are

$$\begin{aligned}
&\lambda_1 \approx 6.78 \cdot 10^{10},\\
&\lambda_2 \approx 49.9,\\
&\lambda_3 \approx 4.23,\\
&\lambda_4 \approx 1.11, \text{ and}\\
&\lambda_5 \approx -0.268.
\end{aligned}$$

Indeed, the number of positive eigenvalues is 4, while the number of negative eigenvalues is 1.

The rank of Q_1 is 5, so there are 5 non-zero elements in $V(I)$. ◇

8 The class of slow-fast systems

Next, another subclass of polynomial dynamical system will be studied: so called *slow-fast polynomial dynamical systems.*

8.1 Introduction

A dynamical system of the form

$$\begin{cases}
\dot{x_1} = & F_1(x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_m; \epsilon)\\
\dot{x_2} = & F_2(x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_m; \epsilon)\\
\vdots & \\
\dot{x_n} = & F_n(x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_m; \epsilon)\\
\epsilon\dot{y_1} = & G_1(x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_m; \epsilon)\\
\epsilon\dot{y_2} = & G_2(x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_m; \epsilon)\\
\vdots & \\
\epsilon\dot{y_m} = & G_m(x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_m; \epsilon)
\end{cases}, \tag{8.1}$$

where ϵ is a parameter such that $0 < \epsilon \ll 1$, and $F_i, G_j : \mathbb{R}^{n+m} \to \mathbb{R}$, is called a slow-fast system.

Remark *The term slow-fast system is used in, for example, [3]. Another term for the same type of systems is singularly perturbed systems; see e.g. [16, chapter 1.3].*

Definition 8.1.1. *If* (8.1) *is also a polynomial dynamical system, we say that it is a polynomial slow-fast system.*

As was remarked above, another term for slow-fast systems is *singularly perturbed systems.* More specifically, slow-fast systems are singular in the following sense: for $\epsilon = 0$, the system becomes a differential-algebraic system; the latter is a system which consists of both differential equations and algebraic equations.

The system (8.1) depends on ϵ; therefore, so will the solutions. We will use two different notations to denote the solution of (8.1) corresponding to a certain choice of ϵ.

Convention *$(x(t;\epsilon), y(t;\epsilon))$ and $(x_\epsilon(t), y_\epsilon(t))$ will both denote the solution of a slow-fast system.*

The second type of notation is natural to use when we surpress t in the notation, which we often do. Since $0 < \epsilon \ll 1$, there should be no risk of the reader interpreting x_ϵ or y_ϵ as a component of the vector x or y, respectively, even when $\epsilon = 0$.

8.2 Application of the theory of polynomial dynamical systems

We can use the method presented in Section 3 to find the steady states of a polynomial slow-fast system.

Example 8.2.1. In [6, Chapter 2.6], the *FitzHugh-Nagumo system* is defined as

$$\begin{cases} \dot{x} = & -\gamma x + \qquad\qquad y \\ \epsilon\dot{y} = & -Cx + \quad Ay(y-\beta)(\delta-y) \end{cases} \tag{8.2}$$

where $\beta, \gamma, \delta, C \in \mathbb{R}$. If we assume that $0 < \epsilon \ll 1$, this is a polynomial slow-fast system.

The ideal

$$I = \left\langle -\gamma x + y,\ -\frac{C}{\epsilon}x + \frac{A}{\epsilon}y(y-\beta)(\delta-y) \right\rangle$$

has a Gröbner basis

$$\begin{aligned} \{ & -A\gamma^3x^3 + A\gamma^2(\beta+\delta)x^2 - (A\beta\gamma\delta + C)x, \\ & y - \gamma x \} \end{aligned}$$

with respect to the Lex-ordering with $y > x$. Thus, $\ell\,(I) = \langle x^3, y \rangle$. This implies, by Proposition 3.8.1, that $\mathcal{B} = \{1 + I,\ x + I,\ x^2 + I\}$ is a basis of $\mathbb{C}[x,y]/I$.

Let

$$\alpha_1 = \frac{\beta+\delta}{\gamma}$$

and

$$\alpha_2 = \frac{A\beta\gamma\delta + C}{A\gamma^3}.$$

In $C[x,y]/I$, we have $x^3 + I = \alpha_1 x^2 - \alpha_2 x + I$. Let T_p denote the linear transformation on $\mathbb{R}[x]/I$ induced by $p \in \mathbb{R}[x]$. Then, in $\mathcal{B}$,

$$[T_1]_{\mathcal{B}} = I$$

and

$$[T_x]_{\mathcal{B}} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & -\alpha_2 \\ 0 & 1 & -\alpha_1 \end{pmatrix}.$$

Note that

$$\begin{aligned} T_{x^n}(f+I) &= x^n f + I \\ &= x^{n-1}(xf) + I \\ &= x^{n-1} T_x(f+I) \\ &= (T_{x^{n-1}} \circ T_x)(f+I) \end{aligned}$$

so $T_{x^n} = (T_x)^n$. This gives

$$\begin{aligned} [T_{x^2}]_{\mathcal{B}} &= \left([T_x]_{\mathcal{B}}\right)^2 \\ &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & -\alpha_2 & \alpha_1\alpha_2 \\ 1 & -\alpha_1 & \alpha_1^2 - \alpha_2 \end{pmatrix}, \end{aligned}$$

$$\begin{aligned} [T_{x^3}]_{\mathcal{B}} &= \left([T_x]_{\mathcal{B}}\right)^3 \\ &= \begin{pmatrix} 0 & 0 & 0 \\ -\alpha_2 & \alpha_1\alpha_2 & -(\alpha_1^2 - \alpha_2)\alpha_2 \\ \alpha_1 & \alpha_1^2 - \alpha_2 & -\alpha_1(\alpha_1^2 - \alpha_2) + \alpha_1\alpha_2 \end{pmatrix}, \text{ and} \end{aligned}$$

$$\begin{aligned} [T_{x_4}]_{\mathcal{B}} &= \left([T_x]_{\mathcal{B}}\right)^4 \\ &= \begin{pmatrix} 0 & 0 & 0 \\ \alpha_1\alpha_2 & -\alpha_1^2\alpha_2 + \alpha_2^2 & \alpha_1(\alpha_1^2 - \alpha_2)\alpha_2 - \alpha_1\alpha_2^2 \\ \alpha_1^2 - \alpha_2 & -\alpha_1(\alpha_1^2 - \alpha_2) + \alpha_1\alpha_2 & (\alpha_1^2 - \alpha_2)^2 - \alpha_1^2\alpha_2 \end{pmatrix}. \end{aligned}$$

We compute the trace of each transformation; this gives

$$\begin{aligned} \operatorname{tr}(T_1) &= 3, \\ \operatorname{tr}(T_x) &= -\alpha_1, \\ \operatorname{tr}(T_{x^2}) &= \alpha_1^2 - 2\alpha_2, \\ \operatorname{tr}(T_{x^3}) &= 2\alpha_1\alpha_2 - \alpha_1(\alpha_1^2 - \alpha_2), \text{ and} \\ \operatorname{tr}(T_{x^4}) &= \alpha_2^2 - 2\alpha_1^2\alpha_2 + (\alpha_1^2 - \alpha_2)^2. \end{aligned}$$

Thus,

$$Q_1 = \begin{pmatrix} 3 & -\alpha_1 & \alpha_1^2 - 2\alpha_2 \\ -\alpha_1 & \alpha_1^2 - 2\alpha_2 & 2\alpha_1\alpha_2 - \alpha_1(\alpha_1^2 - \alpha_2) \\ \alpha_1^2 - 2\alpha_2 & 2\alpha_1\alpha_2 - \alpha_1(\alpha_1^2 - \alpha_2) & \alpha_2^2 - 2\alpha_1^2\alpha_2 + (\alpha_1^2 - \alpha_2)^2 \end{pmatrix}.$$

Q_1 has the characteristic polyomial

$$\begin{aligned} \chi_{Q_1}(t) = \quad & \alpha_1^2\alpha_2^2 - 4\alpha_2^3 \\ & + (-2\alpha_1^2 - 2\alpha_1^4 + 6\alpha_2 + 8\alpha_1^2\alpha_2 - 2\alpha_2^2 - \alpha_1^2\alpha_2^2 + 4\alpha_2^3)t \\ & + (3 + \alpha_1^2 + \alpha_1^4 - 2\alpha_2 - 4\alpha_1^2\alpha_2 + 2\alpha_2^2)t^2 \\ & - t^3. \end{aligned}$$

Recall from linear algebra that the determinant of a matrix is given by the constant term in its characteristic polynomial. Thus,

$$\begin{aligned}\det Q_1 &= \alpha_1^2\alpha_2^2 - 4\alpha_2^3\\ &= \alpha_2^2\left(\alpha_1^2 - 4\alpha_2\right).\end{aligned}$$

Since β, γ, δ, A and C are all positive, so are α_1 and α_2. Thus, $\det Q_1 = 0$ if and only if $\alpha_1^2 = 4\alpha_2$. In other words, Q_1 has rank 3 if and only if $\alpha_1^2 \neq 4\alpha_2$. Assume $\alpha_1^2 = 4\alpha_2$. Then

$$Q_1 = \begin{pmatrix} 3 & -\alpha_1 & \frac{\alpha_1^2}{2} \\ -\alpha_1 & \frac{\alpha_1^2}{2} & -\frac{\alpha_1^3}{4} \\ \frac{\alpha_1^2}{2} & -\frac{\alpha_1^3}{4} & \frac{\alpha_1^4}{8} \end{pmatrix}$$

which has rank 2. Thus,

$$|V(I)| = \begin{cases} 2, & \text{if } \alpha_1^2 = 4\alpha_2 \\ 3, & \text{otherwise} \end{cases}.$$

The number of positive eigenvalues of Q_1 are given by the number of sign changes, denoted k_+, in the sequence of coefficients of the characteristic polynomial, i.e the number of sign changes in the sequence

$$\begin{aligned}&(-1,\\ &3+\alpha_1^2+\alpha_1^4-2\alpha_2-4\alpha_1^2\alpha_2+2\alpha_2^2,\\ &-2\alpha_1^2-2\alpha_1^4+6\alpha_2+8\alpha_1^2\alpha_2-2\alpha_2^2-\alpha_1^2\alpha_2^2+4\alpha_2^3,\\ &\alpha_1^2\alpha_2^2-4\alpha_2^3),\end{aligned}$$

while the number of negative eigenvalues are given by the number of sign changes, denoted k_-, in

$$\begin{aligned}&(1,\\ &3+\alpha_1^2+\alpha_1^4-2\alpha_2-4\alpha_1^2\alpha_2+2\alpha_2^2,\\ &-(-2\alpha_1^2-2\alpha_1^4+6\alpha_2+8\alpha_1^2\alpha_2-2\alpha_2^2-\alpha_1^2\alpha_2^2+4\alpha_2^3),\\ &\alpha_1^2\alpha_2^2-4\alpha_2^3).\end{aligned}$$

Then the signature of Q_1 can be computed, since $\operatorname{sign}(Q_1) = k_+ - k_-$, by the definition of signature.

For example, let

$$\beta = 1, \quad \gamma = 1, \quad \delta = 1,$$

$$A = \frac{1}{6}, \quad C = 1.$$

Then $\alpha_1 = \alpha_2 = 1$. Then k_+ is the number of sign changes in

$$(-1, 1, 11, -3),$$

so $k_+ = 2$, and k_- is the number of sign changes in

$$(1, 1, -11, -3)\,,$$

so $k_- = 1$. This gives $\operatorname{sign}(Q_1) = 1$; hence, in this case, the FitzHugh-Nagumo system has one steady state.

We can find the steady state using the algorithm presented earlier in the thesis. The solutions of the equation

$$\begin{aligned} & -A\gamma^3 x^3 + A\gamma^2(\beta+\delta)x^2 - (A\beta\gamma\delta + C)x = 0 \\ \Leftrightarrow & x^3 - \alpha_1 x^2 + \alpha_2 x = 0 \end{aligned}$$

are

$$x = 0, \text{ and}$$

$$x = \frac{\alpha_1}{2} \pm \sqrt{\frac{\alpha_1^2 - 4\alpha_2}{4}}.$$

Substituting this into $-\gamma x + y = 0$, we can solve for y as well. This gives that the steady states of the FitzHugh-Nagumo system are

- just $(0, 0)$ if $\alpha_1^2 < 4\alpha_2$,
- the points $(0, 0)$ and $\left(\frac{\alpha_1}{2}, \frac{\gamma\alpha_1}{2}\right)$ if $\alpha_1^2 = 4\alpha_2$, and
- the points $(0, 0)$,

$$\left(\frac{\alpha_1 + \sqrt{\alpha_1^2 - 4\alpha_2}}{2}, \gamma\frac{\alpha_1 + \sqrt{\alpha_1^2 - 4\alpha_2}}{2}\right)$$

  and

$$\left(\frac{\alpha_1 - \sqrt{\alpha_1^2 - 4\alpha_2}}{2}, \gamma\frac{\alpha_1 - \sqrt{\alpha_1^2 - 4\alpha_2}}{2}\right)$$

  if $\alpha_1^2 > 4\alpha_2$.

Let us check that this is in agreement with what the signature of Q_1 tells us. If $\alpha_1^2 = 4\alpha_2$, then

$$\chi_{Q_1}(t) = -t^3 + \frac{1}{64}\left(192 + 32\alpha_1^2 + 8\alpha_1^4\right)t^2 - \frac{1}{64}\left(32\alpha_1^2 + 8\alpha_1^4\right).$$

Since $192 + 32\alpha_1^2 + 8\alpha_1^4 > 0$ and $32\alpha_1^2 + 8\alpha_1^4 > 0$, this gives $k_+ = 2$. Substituting t for $-t$ gives that $k_- = 0$. Thus, the number of steady states is two, which is the number of steady states we found. We can proceed in the same way for the other cases, but they are messier, so to simplify matters, let us check this for just one choice of (α_1, α_2) per case. Above, we considered the case $\alpha_1 = \alpha_2 = 1$, which falls into the case $\alpha_1^2 < 4\alpha_2$, and we saw the the number of steady states should be one, which is in agreement with that the only steady state of the system is $(0, 0)$. A choice which falls into the case $\alpha_1 > 4\alpha_2$ is $\alpha_1 = 3$ and $\alpha_2 = 1$. In this case, k_+ is the number of sign changes in the sequence

$$(-1, 57, -109, 5)$$

and k_- is the number of sign changes in the sequence

$$(1, 57, 109, 5).$$

Thus, the number of steady states is $k_+ - k_- = 3 - 0 = 3$, which is is the number of steady states which we found above. ◇

9 The class of homogeneous polynomial dynamical systems

Proposition 9.0.1. *Consider the system $\dot{x}_i = p_i(x)$, $i = 1, 2, \ldots, n$, where each p_i is a homogeneous polyomial. Let $P = \{p_1, p_2, \ldots, p_n\}$. Let $\Delta = \{\deg(p) \mid p \in P\}$ and index the elements $\delta_i \in \Delta$ so that $\delta_1 > \delta_2 > \cdots > \delta_m$. Let $k_i = |\{p \in P | \deg(p) = i\}|$. Let $s_j = \sum_{i=1}^{j} k_j$. Then the system can be written on the form*

$$\begin{cases} \dot{x}_1 = & p_1(x) \\ \dot{x}_2 = & p_2(x) \\ \vdots & \\ \dot{x}_{s_1} = & p_{s_1}(x) \\ \lambda^{\delta_1 - \delta_2} \dot{x}_{s_1+1} = & p_{s_1+1}(x) \\ \vdots & \\ \lambda^{\delta_1 - \delta_2} \dot{x}_{s_2} = & p_{s_2}(x) \\ \vdots & \\ \lambda^{\delta_1 - \delta_m} \dot{x}_{s_{m-1}+1} = & p_{s_{m-1}+1}(x) \\ \vdots & \\ \lambda^{\delta_1 - \delta_m} \dot{x}_n = & p_n(x) \end{cases}$$

where λ is an arbitrary parameter and $\deg(p_i) = \delta_j$ *for $s_{j-1} < i \leq s_j$.*

Proof. Let

$$\tilde{x_i}(\tau) = \frac{1}{\lambda} x_i(\tau \lambda^{-(\delta_1 - 1)}).$$

This gives

$$\frac{d\tilde{x_i}}{d\tau} = \frac{1}{\lambda^{\delta_1}} \left. \frac{dx_i}{dt} \right|_{t = \tau \lambda^{-(\delta_1 - 1)}}.$$

Assume that $\deg(p_i) = \delta_k$. Then

$$\frac{d\tilde{x_i}}{d\tau} = \frac{1}{\lambda^{\delta_1}} \lambda^{\delta_k} p_i(\tilde{x}) = \lambda^{\delta_k - \delta_1} p_i(\tilde{x}),$$

since p_i is homogeneous of degree d_i. Thus,

$$\lambda^{\delta_1 - \delta_k} \frac{d\tilde{x_i}}{d\tau} = p_i(\tilde{x}).$$

After renaming of $\tilde{x_i}$ to x_i and τ to t, and re-indexing the x_i and p_i, the conclusion follows. □

10 Conclusion and further work

We have presented a framework for studying important properties of polynomial dynamical systems. In addition to methods for computing the number of and determining the steady states of such systems, we have presented methods to reduce the dimension and the number of parameters of a system. We have also shown how the framework can be applied to some subclasses of polynomial dynamical systems.

One line of further work is to try to find a general method which, given a certain subclass of polynomial dynamical systems, can give a characterization of the subclass of systems in terms of its parameters, e.g. for which parameters does a certain subclass of systems have m steady states, for which parameters are these stable/unstable/asymptotically stable, and so on. While the methods presented in this thesis is very useful for studying polynomial dynamical systems one at a time, it would be of course even more useful to deal with a whole class of systems at once. Could we not use the methods presented in this thesis on systems with unknown parameters (i.e. on a whole class of system) and then check for which ranges of parameters the system has a certain property? Yes, we could, using ad hoc methods — but we would like to have a general method, which we can apply to any given subclass of polynomial dynamical systems. To summarize in different terms: we would like to find a general method for studying the parameter space of subclasses of systems.

A Appendix

Proposition A.0.1. *Let A be an $m \times n$-matrix. Let E be a non-singular $m \times m$-matrix. Then* $\ker EA = \ker A$.

Proof. On the one hand,

$$\begin{aligned} & v \in \ker A \\ \Leftrightarrow\quad & Av = 0 \\ \Rightarrow\quad & EAv = 0 \\ \Leftrightarrow\quad & v \in \ker EA, \end{aligned}$$

so $\ker A \subset \ker EA$, while on the other hand,

$$\begin{aligned} & v \in \ker EA \\ \Leftrightarrow\quad & EAv = 0 \\ \Rightarrow\quad & Av = E^{-1}EAv = 0, \text{ since } E \text{ is non-singular} \\ \Leftrightarrow\quad & v \in \ker A, \end{aligned}$$

so $\ker EA \subset \ker A$. Hence, $\ker EA = \ker A$. □

Proposition A.0.2. *T_p, defined in Definition 5.1.11, is a linear tranformation.*

Proof. Let $f + I, g + I \in k[x]/I$ and $\alpha, \beta \in k$. Then

$$\begin{aligned} T_p(\alpha(f+I) + \beta(g+I)) &= T_p((\alpha f + \beta g) + I) \\ &= (p\alpha f + p\beta g) + I \\ &= \alpha p(f+I) + \beta p(g+I) \\ &= \alpha T_p(f) + \beta T_p(g). \end{aligned}$$

□

Proposition A.0.3. *B_q, defined in Definition 5.1.12, is a symmetric bilinear form.*

Proof. Let $p_1, p_2, q_1, q_2 \in k[x]$ and let $\alpha_1, \alpha_2, \beta_1, \beta_2 \in k$. Then

$$\begin{aligned} & B_q(\alpha_1 p_1 + \alpha_2 p_2, \beta_1 q_1 + \beta_2 q_2) \\ = \; & \operatorname{tr}\left(T_{q(\alpha_1 p_1 + \alpha_2 p_2)(\beta_1 q_1 + \beta_2 q_2)}\right) \\ = \; & \operatorname{tr}\left(\alpha_1\beta_1 T_{qp_1q_1} + \alpha_1\beta_2 T_{qp_1q_2} + \alpha_2\beta_1 T_{qp_2q_1} + \alpha_2\beta_2 T_{qp_2q_2}\right) \\ = \; & \alpha_1\beta_1 \operatorname{tr}\left(T_{qp_1q_1}\right) + \alpha_1\beta_2 \operatorname{tr}\left(T_{qp_1q_2}\right) + \alpha_2\beta_1 \operatorname{tr}\left(T_{qp_2q_1}\right) + \alpha_2\beta_2 \operatorname{tr}\left(T_{qp_2q_2}\right) \\ = \; & \alpha_1\beta_1 B_q(p_1, q_1) + \alpha_1\beta_2 B_q(p_1, q_2) + \alpha_2\beta_1 B_q(p_2, q_1) + \alpha_2\beta_2 B_q(p_2, q_2) \end{aligned}$$

where

$$\begin{aligned} & T_{\alpha p + \beta q}(f + I) = \alpha p(f + I) + \beta q(f + I) = \left(\alpha T_p + \beta T_q\right)(f + I), \text{ and} \\ & \operatorname{tr}\left(\alpha_1 T_1 + \alpha_2 T_2\right) = \alpha_1 \operatorname{tr}\left(T_1\right) + \alpha_2 \operatorname{tr}\left(T_2\right) \end{aligned}$$

have been used. Also,

$$B_q(p_1, p_2) = \operatorname{tr}\left(T_{qp_1p_2}\right) = \operatorname{tr}\left(T_{qp_2p_1}\right) = B_q(p_2, p_1),$$

so it is symmetric. □

References

[1] Michael F. Atiyah and Ian G. Macdonald. *Introduction to commutative algebra.* Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont., 1969.

[2] Saugata Basu, Richard Pollack, and Marie-Françoise Roy. *Algorithms in Real Algebraic Geometry.* Algorithms and Computation in Mathematics. Springer Berlin Heidelberg, 2013.

[3] Nils Berglund and Barbara Gentz. *Noise-induced phenomena in slow-fast dynamical systems.* Probability and its Applications (New York). Springer-Verlag London, Ltd., London, 2006. A sample-paths approach.

[4] David A. Cox, John Little, and Donal O'Shea. *Ideals, varieties, and algorithms.* Undergraduate Texts in Mathematics. Springer, Cham, fourth edition, 2015. An introduction to computational algebraic geometry and commutative algebra.

[5] Leah Edelstein-Keshet. *Mathematical models in biology.* The Random House/Birkhäuser Mathematics Series. Random House, Inc., New York, 1988.

[6] Christopher P. Fall, Eric S. Marland, John M. Wagner, and John J. Tyson, editors. *Computational cell biology*, volume 20 of *Interdisciplinary Applied Mathematics.* Springer-Verlag, New York, 2002.

[7] Ralf Fröberg. *An introduction to Gröobner bases.* Pure and Applied Mathematics (New York). John Wiley & Sons, Ltd., Chichester, 1997.

[8] Paul A. Fuhrmann. *A polynomial approach to linear algebra*. Universitext. Springer, New York, second edition, 2012.

[9] F.R. Gantmacher. *The Theory of Matrices*. AMS Chelsea Publishing, 1957.

[10] M.W. Hirsch, R.L. Devaney, and S. Smale. *Differential Equations, Dynamical Systems, and Linear Algebra*. Pure and Applied Mathematics. Elsevier Science, 1974.

[11] Bhubaneswar Mishra. *Algorithmic algebra*. Texts and Monographs in Computer Science. Springer-Verlag, New York, 1993.

[12] Paul Pedersen, Marie-Françoise Roy, and Aviva Szpirglas. Counting real zeros in the multivariate case. In *Computational algebraic geometry*, pages 203–224. Springer, 1993.

[13] Lawrence Perko. *Differential equations and dynamical systems*, volume 7 of *Texts in Applied Mathematics*. Springer-Verlag, New York, 1991.

[14] M. Reid. *Undergraduate Commutative Algebra*. London Mathematical Society St. Cambridge University Press, 1995.

[15] W. Rudin. *Principles of Mathematical Analysis*. International series in pure and applied mathematics. McGraw-Hill, 1976.

[16] Elena Shchepakina, Vladimir Sobolev, and Michael P. Mortell. *Singular Perturbations: Introduction to System Order Reduction Methods with Applications*. Lecture Notes in Mathematics. Springer International Publishing, 2014.

[17] Eduardo D. Sontag. *Lecture notes on Mathematical Systems Biology*. 2018. Version 7.7.5.

[18] Gerald Teschl. *Ordinary differential equations and dynamical systems*, volume 140 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2012.