# Ideals in Numerical Applications

Tomas Sauer

Lehrstuhl für Mathematik mit Schwerpunkt Digitale Bildverarbeitung
FORWISS
University of Passau
Innstr. 43
94032 Passau


Version 0.0
14.9.2019

*Nothing spoils numbers faster than a lot of arithmetic.*

Peppermint Patty, *The Peanuts, 4.12.1968*

*Of course she was aware, cognitively, that there was a life outside universities, but she knew nothing about it,*

D. Lodge, *Nice Work*

*To isolate mathematics from the practical demands of the sciences is to invite the sterility of a cow shut away from the bulls.*

P. Chebychev

*. . . you get to have such a high regard for the truth you can't put courtesy first. You want to, but you haven't the heart.*

E. D. Biggers, *Charlie Chan . . .*

*Reality is software. What does it matter what system it's running on?*

R. Rucker, *Postsingular*

*And thus there seems a reason in all things, even in law.*

H. Melville, *Moby Dick*

*What the eye does not see, the stomach does not get upset over.*

J. K. Jerome, *Three Men in a Boat*

# Contents

# 1 The univariate case

*If you've got it all and you're still unhappy, what's the point of everything?*

(I. Rankin, *Dead souls*)

In this first chapter, we give a quick overview over the concepts we are going to consider in this lecture and how they look in the univariate case. We will see in many instances that the univariate case is extraordinarily simple and most of the lecture will deal with the problem how we can extend the ideas and concepts step by step. Nevertheless, this chapter may serve as a guideline and help to motivate the long way through the jungle of more sophisticated algebraic concepts. To that end, we will sometimes consider a slightly eccentric perspective of the respective problem, but the simple reason for that is the better compatibility with the multivariate situation.

## 1.1 Polynomial basics

Polynomials are among the most classical and useful concepts in mathematics. As functions, they can be represented by finitely many coefficients, hence stored and manipulated on a computer.

**Definition 1.1.1** (Polynomials & Laurent polynomials)**.** The ring of polynomials $\Pi = \mathbb{K}[x]$ in one variables with coefficients in the field $\mathbb{K}$ is defined as the set of all finite sums of powers of $x$ equipped with coefficients in $\mathbb{K}$. Therefore, a POLYNOMIAL is an expression of the form

$$f(x) = \sum_{k\in\mathbb{N}_0} f_k\, x^k, \qquad f_k \in \mathbb{K}, \qquad \#\{k : f_k \neq 0\} < \infty. \tag{1.1.1}$$

In the same way, a LAURENT POLYNOMIAL is of the form

$$f(x) = \sum_{k\in\mathbb{Z}} f_k\, x^k, \qquad f_k \in \mathbb{K}, \qquad \#\{k : f_k \neq 0\} < \infty, \tag{1.1.2}$$

and the ring of Laurent polynomials will be denoted by $\Lambda$.

The difference between polynomials and Laurent polynomials is that the latter also admit negative powers of $x$ and thus are not defined at $x = 0$. On the other hand, for any Laurent polynomial $f \in \Lambda$ there exist $k \in \mathbb{Z}$ and $p \in \Pi$ such that $f(x) = x^{-k}\, p(x)$, and we can even normalize $k$ such that $p(0) \neq 0$, i.e., $p_0 \neq 0$, as long as $f \neq 0$. The monomials are units in the ring $\Lambda$, $\left(x^k\right)^{-1} = x^{-k}$, $k \in \mathbb{Z}$, so it seems that the modification from polynomials to Laurent polynomials is a very minor one. This is not true, we will see later that the two rings have a totally different structure and that this has consequences.

**Exercise 1.1.1** Determine the set of units in $\Pi$ and $\Lambda$. Recall that $a \in R$ is called a UNIT in the ring $R$ if there exists $a^{-1} \in R$ such that $a^{-1}a = 1$. For simplicity we only consider commutative rings with unit element 1 here. ♢

**Remark 1.1.2.** Sometimes Laurent polynomials are also introduced by writing $y$ for $x^{-1}$ and then consider the *bivariate* Polynomials $\mathbb{K}[x,y]$ with the additional requirement that $xy = 1$. Using ideal notation that we will introduce in more detail later, we can then write this as $\Lambda = \mathbb{K}[x,y]/\langle xy-1\rangle$. This is called a LOCAL RING, cf. [Eisenbud, 1994].

**Definition 1.1.3** (Degree & leading term)**.** Let $f \in \Pi$ be a polynomial.

1. The DEGREE of $f$ is defined as
$$\deg f = \max\{k : f_k \neq 0\}, \tag{1.1.3}$$
with the convention that the degree of the zero polynomial is $-1$.
2. The LEADING TERM of $f$ is
$$\lambda(f) := f_{\deg f}\, x^{\deg f} \in \Pi, \tag{1.1.4}$$
and the LEADING COEFFICIENT[1] is $\kappa(f) := f_{\deg f} \in \mathbb{K}$.
3. A polynomial is called MONIC if $\kappa(f) = 1$.

In Definition 1.1.3 we introduced the degree and the related concepts only for polynomials, not for Laurent polynomials. This has a simple reason: there is no notion of degree for Laurent polynomials, not even in the much more general multivariate context of a graded ring that we will consider later.

The degree is a good measure for the complexity of a polynomial. In general, polynomials get more complicated, oscillatory and misbehaving if the degree increases.

### 1.1.1 Division with remainder

Division with remainder, also called POLYNOMIAL DIVISION or LONG DIVISION is a standard procedure in elementary algebra. Given $f, g \in \Pi$ it computes a decomposition

$$f = q\,g + r, \qquad q, r \in \Pi, \qquad \deg r < \deg g, \tag{1.1.5}$$

where the polynomial $r$ is called the REMAINDER of the division, written as $(f)_g := r$. This is a common property between polynomials and integers and makes both of them examples of a EUCLIDEAN RING, cf. [Gathen and Gerhard, 1999]. Let us recall the algorithm.

**Algorithm 1.1.1** Division with remainder: $f, g \in \Pi$, $g \neq 0$

$p \leftarrow f$
$q \leftarrow 0$
**while** $\deg p \geq \deg g$ **do**
    $q \leftarrow q + \dfrac{\lambda(p)}{\lambda(g)}$
    $p \leftarrow p - \dfrac{\lambda(p)}{\lambda(g)}\, g$
**end while**
$r \leftarrow p$

[1] "$\kappa$" like "leading $\kappa$oefficient".

**Remark 1.1.4.** The division algorithm is particularly simple if the polynomial $g$ is monic as then the division by $\lambda(g)$ only means a shift of the exponent of the monomial. Therefore, the divisor polynomial is sometimes normalized before division which is critical if $\lambda_{\mathbb{K}}(g)$ is small relative to the other coefficient which in turn means that the "true" degree of $g$ is smaller than $\deg g$. The problem with almost zero floating point numbers is well–known and discussed a lot in the literature, cf. [Higham, 2002].

The procedure is indeed simple: We subtract *polynomial* multiples of $g$ from $p$ in such a way that the leading terms of the two polynomials in the subtraction coincide, thus reducing the degree by 1 in each step. Since the degree is finite, this procedure terminates after finitely many steps and leaves the remainder $r$. Collecting the factors in each step gives $q$. Therefore, Algorithm 1.1.1 computes the decomposition (1.1.5). In particular, $q$ and $r$ are *unique*.

**Exercise 1.1.2** Prove the validity of Algorithm 1.1.1 and the uniqueness of $q$ and $r$. ♢

Even if it is a triviality, let us remark it here: *$f$ is a multiple of $g$ if and only if $(f)_g = 0$ in* (1.1.5). We mention this property because its multivariate analogue will be the basis for Gröbner basis constructions. Let us this for a somewhat strange definition which, on the other hand, can be transferred to the multivariate case directly.

**Definition 1.1.5.** $f \in \Pi$ is called DIVISIBLE by $g \in \Pi$ if $(f)_g = 0$. We write this as $g|f$ and say that *"$g$ divides $f$"*.

### 1.1.2 Euclidean algorithm, greatest common divisors and principal ideals

Having defined divisibility, we can start to talk about common divisors and, of course, the greatest among them.

**Definition 1.1.6.**

1. $g \in \Pi$ is called a DIVISOR of $f \in \Pi$ if $g|f$.
2. $g \in \Pi$ is called a COMMON DIVISOR of $f_1, \ldots, f_n \in \Pi$ if $g|f_j$, $j = 1, \ldots, n$.
3. $g \in \Pi$ is called a GREATEST COMMON DIVISOR of $f_1, f_2 \in \Pi$, written as $g = \gcd(f_1, f_2)$ if $g$ is a common divisor of $f_1, f_2$ and whenever $h$ is a common divisor of $f_1, f_2$ it follows that $h|g$.

The greatest common divisor is *not* unique for polynomials. Indeed, $c\gcd(f_1, f_2)$, $c \in \mathbb{K} \setminus \{0\}$, is another greatest common divisor, and all gcds are of this form. In fact, in any ring any unit multiple of $g = \gcd(f_1, f_2)$ is a gcd again: if $a$ is a unit in $R$ then

$$f_j = q\,g = \left(a^{-1}q\right)(ag), \qquad j = 1, 2,$$

hence $ag$ is a common divisor as well that is divided by any other common divisor, which is shown in exactly the same way. If we would need it, we could make **the** common divisor unique by choosing the monic gcd, i.e., selecting **the** greatest common divisor $\gcd^*$ as

$$\gcd^*(f_1, f_2) = x^n + \cdots = \frac{\gcd(f_1, f_2)}{\kappa\left(\gcd(f_1, f_2)\right)},$$

cf. [Gathen and Gerhard, 1999]. But this is neither necessary nor overly useful, so we prefer to live with the ambiguity. The computation of the gcd is done by the classical EUCLIDEAN ALGORITHM based on iterated division with remainder. We give a slightly more advanced version of this algorithm here, namely the EXTENDED EUCLIDEAN ALGORITHM in Algorithm 1.1.2. The standard algorithm is obtained by considering the $r_j$ only.

**Algorithm 1.1.2** EXTENDED EUCLIDEAN ALGORITHM: $f, g \in \Pi \setminus \{0\}$

$r_0 \leftarrow f, \quad p_0 \leftarrow 1, \quad q_0 \leftarrow 0$
$r_1 \leftarrow g, \quad p_1 \leftarrow 0, \quad q_1 \leftarrow 1$
$j \leftarrow 1$
**while** $r_j \neq 0$ **do**
    $r_{j-1} = s_j\, r_j + r_{j+1}$ (Division with remainder, defines $s_{j-1}, r_{j+1}$)
    $p_{j+1} \leftarrow p_{j-1} - s_j\, p_j$
    $q_{j+1} \leftarrow q_{j-1} - s_j\, q_j$
    $j \leftarrow j + 1$
**end while**
$\gcd(f, g) \leftarrow r_{j-1}$
$p \leftarrow p_{j-1}, \quad q \leftarrow q_{j-1}$

**Theorem 1.1.7.** *The extended euclidean algorithm*

1. *terminates after finitely many steps,*
2. *computes* $\gcd(f, g)$,
3. *computes* BÉZOUT COEFFICIENTS $p, q \in \Pi$ *such that*

$$\gcd(f, g) = f\, p + g\, q. \tag{1.1.6}$$

**Proof:** We write the relation between the $r_j$ as

$$r_{j+1} = \left(r_{j-1}\right)_{r_j}, \qquad j \in \mathbb{N}, \qquad r_0 = f, \quad r_1 = g,$$

and since $\deg r_{j+1} < \deg r_j$, the algorithm has to terminate after finitely many steps. By the iteration

$$r_{j+1} = r_{j-1} - s_j\, r_j, \qquad j \in \mathbb{N}, \tag{1.1.7}$$

and $r_0 = f$, $r_1 = g$, it follows inductively that $\gcd(f, g) | r_j$, $j \in \mathbb{N}_0$. Choose $n$ such that $r_{n+1} = 0 \neq r_n$, then (1.1.7) with $j = n$ yields that $r_n | r_{n-1}$, and because of the backwards iteration

$$r_{j-1} = s_j r_j + r_{j+1}, \qquad j = n-1, \ldots, 1,$$

$r_n$ also divides $r_{n-2}, r_{n-3}, \ldots, r_1 = g, r_0 = f$, hence $r_n = \gcd(f, g)$ as claimed in 2). (1.1.6) is the case $j = n$ of the invariance

$$p_j\, f + q_j\, g = r_j, \qquad j = 0, \ldots, n, \tag{1.1.8}$$

which we prove by induction. The initialization ensures the validity of (1.1.8) for $j = 0, 1$ while for $j \geq 1$ we have that

$$\begin{aligned} p_{j+1}\, f + q_{j+1}\, g &= \left(p_{j-1} - s_j p_j\right) f + \left(q_{j-1} - s_j q_j\right) g = \left(p_{j-1} f + q_{j-1} g\right) - s_j (p_j f + q_j g) \\ &= r_{j-1} - s_j\, r_j = r_{j+1}, \end{aligned}$$

which completes the proof. □

The formulation of the extended euclidean algorithm in Algorithm 1.1.2 with all its indices is inefficient and was only for the purpose of the proof of Theorem 1.1.7. The more appropriate version is by means of a matrix.

**Algorithm 1.1.3** Matrix version of extended euclidean algorithm: $f, g \in \Pi \setminus \{0\}$

1: $R \leftarrow \begin{pmatrix} f & 1 & 0 \\ g & 0 & 1 \end{pmatrix} = \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \end{pmatrix}$
2: $r \leftarrow (r_{11})_{r_{21}}$
3: **while** $r \neq 0$ **do**
4: $\quad s \leftarrow r_{11}/r_{21}$
5: $\quad R \leftarrow \begin{pmatrix} r_{21} & r_{22} & r_{23} \\ r & r_{12} - s\,r_{22} & r_{13} - s\,r_{23} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & -s \end{pmatrix} R$
6: $\quad r \leftarrow (r_{11})_{r_{21}}$
7: **end while**

**Example 1.1.8.** Let us illustrate the algorithm by computing the gcd of

$$f = x^3 + 2x^2 + x, \qquad g = x^2 - 1.$$

The matrix in Algorithm 1.1.2 then is computed as follows

$$\begin{pmatrix} x^3 + 2x^2 + x & 1 & 0 \\ x^2 - 1 & 0 & 1 \end{pmatrix} \qquad s = x + 2$$

$$\downarrow$$

$$\begin{pmatrix} x^2 - 1 & 0 & 1 \\ 2x + 2 & 1 & -x - 2 \end{pmatrix}$$

and since $2x + 2 = 2(x + 1) | x^2 - 1$ the algorithm already terminates. The Bézout identity takes the form

$$\left(x^3 + 2x^2 + x\right) - (x + 2)\left(x^2 - 1\right) = 2(x + 1),$$

and if we want the normalized gcd, we simple have to divide both coefficients of the identity by the leading coefficient of the gcd yielding

$$\frac{1}{2}\left(x^3 + 2x^2 + x\right) - \left(\frac{1}{2}x + 1\right)\left(x^2 - 1\right) = x + 1.$$

The gcd computation by means of the extended euclidean algorithm will become the theoretical backbone of this lecture, but unfortunately it has a small but relevant deficit: it does not work in numerical accuracy. Since we care for numerical applications and mostly computations by means of floating point numbers which are contaminated by roundoff errors, or use so called EMPIRICAL POLYNOMIALS whose coefficients are only determined up to a certain accuracy, let us briefly have a look at a numerically stable method of determining the gcd which will already make us acquainted with the spirit of many methods to follows. The approach is taken from [Corless et al., 2004] though in principle even known earlier, cf. [Laidacker, 1969]. For simplicity, here we consider only the case that $f$ and $g$ have simple zeros, multiple zeros make things a little bit more complex.

**Exercise 1.1.3** Show that for any two polynomials $f, g$ and any $\varepsilon > 0$ there exist polynomials of $f_\varepsilon$ and $g_\varepsilon$ of the same degrees whose coefficient are smaller than $\varepsilon$ in absolute value, such that $\gcd\left((f + f_\varepsilon), (g + g_\varepsilon)\right) = 1$. ♢

**Definition 1.1.9.** For $f, g \in \Pi$, $\deg f =: m$, $\deg g =: n$, the SYLVESTER MATRIX is defined as

$$S(f,g) = \begin{pmatrix} f_m & \cdots & f_0 & & \\ & \ddots & \ddots & \ddots & \\ & & f_m & \cdots & f_0 \\ g_n & \cdots & g_0 & & \\ & \ddots & \ddots & \ddots & \\ & & g_n & \cdots & g_0 \end{pmatrix} \in \mathbb{K}^{(m+n)\times(m+n)}, \tag{1.1.9}$$

where the first block of rows is repeated $n$ times, the second one $m$ times.

One well–known property of the Sylvester matrix is that it encodes whether two polynomials are coprime or not.

**Theorem 1.1.10.** $\det S(f,g) \neq 0$ *if and only if* $\gcd(f,g) = 1$.

This result is easily proved by noting that

$$S(f,g)\begin{pmatrix} x^{m+n-1} \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} x^{n-1}f(x) \\ \vdots \\ f(x) \\ x^{m-1}g(x) \\ \vdots \\ g(x) \end{pmatrix}, \tag{1.1.10}$$

hence

$$\begin{pmatrix} x^{m+n-1} \\ \vdots \\ 1 \end{pmatrix} \in \ker S(f,g) \qquad \Leftrightarrow \qquad x \in Z(f,g), \tag{1.1.11}$$

so that $s := \operatorname{rank} S(f,g) \leq n + m - \#Z(f,g)$. Note that even for polynomials with rational or real coefficients, it is not relevant here whether the zero is real or complex. Indeed, if $x$ is a complex common zero of $f$ and $g$, then we can write

$$\begin{pmatrix} x^{m+n-1} \\ \vdots \\ 1 \end{pmatrix} = a + ib, \qquad a, b \in \mathbb{R}^{m+n}$$

and

$$0 = S(f,g)(a+ib) = S(f,g)a + iS(f,g)b \qquad \Rightarrow \qquad 0 = S(f,g)a = S(f,g)b,$$

and the quadratic polynomial $(\cdot - x)(\cdot - \overline{x})$ is a common divisor of $f$ and $g$ connected to this *pair* of kernel elements.

We progress with a $QR$ factorization of the Sylvester matrix,

$$S(f,g) = Q\begin{pmatrix} R & B \\ 0 & 0 \end{pmatrix}, \qquad R \in \mathbb{R}^{s\times s},\ r_{jj} \neq 0, \qquad B \in \mathbb{R}^{s\times(m+n-s)},$$

where $R$ is an upper triangular matrix with nonzero diagonal elements[2]. The polynomial

$$\begin{aligned} h(x) &:= e_s^T \begin{pmatrix} R & B \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x^{m+n-1} \\ \vdots \\ 1 \end{pmatrix} = e_s^T Q^T \begin{pmatrix} x^{n-1} f(x) \\ \vdots \\ f(x) \\ x^{m-1} g(x) \\ \vdots \\ g(x) \end{pmatrix} = e_s^T \begin{pmatrix} p_1(x) \\ \vdots \\ p_{n+m}(x) \end{pmatrix} \gcd(f,g)(x) \\ &= r_{ss} x^{n+m-s} + b_{s1} x^{n+m-s-1} + \cdots + b_{s,n+m-s-1} x + b_{s,n+m-s} \end{aligned}$$

has degree $n+m-s$ and is a multiple of $\gcd(f,g)$, so that $\deg\gcd(f,g) \le n+m-s$. By the Bézout identity (1.1.6) there exist $p, q$ such that $pf + qg = \gcd(f,g)$. We can even show that $\deg f \le n-1$ and $\deg g \le m-1$, see Proposition 1.2.5. Using this fact, we can write $\boldsymbol{p} := (p_{n-1}, \ldots, p_0)$ and $\boldsymbol{q} := (q_{m-1}, \ldots, q_0)$ for the coefficients of $p$ and $q$ and get that

$$\begin{aligned} \gcd(f,g)(x) &= p(x)f(x) + q(x)g(x) = \boldsymbol{p}^T \begin{pmatrix} x^{n-1} \\ \vdots \\ 1 \end{pmatrix} f(x) + \boldsymbol{q}^T \begin{pmatrix} x^{n-1} \\ \vdots \\ 1 \end{pmatrix} g(x) \\ &= (\boldsymbol{p}^T, \boldsymbol{q}^T) \begin{pmatrix} x^{n-1} f(x) \\ \vdots \\ f(x) \\ x^{m-1} g(x) \\ \vdots \\ g(x) \end{pmatrix} = (\boldsymbol{p}^T, \boldsymbol{q}^T) Q \begin{pmatrix} R & B \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x^{n+m-1} \\ \vdots \\ 1 \end{pmatrix} \\ &= \left( Q \begin{pmatrix} \boldsymbol{p} \\ \boldsymbol{q} \end{pmatrix} \right)^T \begin{pmatrix} R & B \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x^{n+m-1} \\ \vdots \\ 1 \end{pmatrix} \end{aligned}$$

hence $\gcd(f,g)$ is a linear combination of the polynomials

$$e_k^T \begin{pmatrix} R & B \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x^{n+m-1} \\ \vdots \\ 1 \end{pmatrix}, \qquad k = 1, \ldots, s$$

each of which is either zero or of degree *exactly* $m+n-k$ since the diagonal elements of $R$ are nonzero. This shows that $\deg h \ge n+m-s$, hence $\deg h = n+m-s$, so that indeed $h$ is a multiple of the gcd.

**Remark 1.1.11.** The value of this approach lies in the fact that it turns an algebraic problem into a problem of **linear algebra**. We will see later in this lecture that the "nonlinearity" of the problem is reflected by the fact that it is turned into an **eigenvalue problem** as determining the kernel of the Sylvester matrix corresponds to computing a structured basis for the eigenspace with respect to the eigenvalue 0. Moreover, and this is the computational aspect, we can rely on techniques from numerical linear algebra which often provides efficient and numerically stable algorithms for such problems. This numerically oriented point of view for the treatment of algebraic problems is fairly recent, see, for example [Stetter, 2005].

[2] By means of proper PIVOTING we can even ensure that the diagonal elements are positive and decreasing, cf. [Golub and van Loan, 1996].

The advantage of the extended euclidean algorithm lies in the definition of the Bézout coefficients and allows us a first simple touch with ideal theory – in its simplest form.

**Definition 1.1.12** (Ideals)**.**

1. A subset $\mathscr{I} \subset \Pi$ of polynomials is called an *ideal* if it is closed under addition and multiplication with arbitrary polynomials, i.e.,
$$f, g \in \mathscr{I}, \quad q \in \Pi \qquad \Rightarrow \qquad f+g \in \mathscr{I}, \quad qf \in \mathscr{I}. \tag{1.1.12}$$
2. The ideal $\langle \mathscr{F} \rangle$ generated by set $\mathscr{F} \subset \Pi$ is the CLOSURE of $\mathscr{F}$ under this operations:
$$\langle \mathscr{F} \rangle = \left\{ \sum_{f \in \mathscr{F}} q_f f : q_f \in \Pi \right\}. \tag{1.1.13}$$
3. An ideal $\mathscr{I} \subset \Pi$ is called a PRINCIPAL IDEAL if it is generated by a single polynomial, that is, there exists $f \in \Pi$ such that $\mathscr{I} = \langle f \rangle$.

Now it is easy to see that any ideal of univariate polynomials is a principal ideal which is the reason why they are called a PRINCIPAL IDEAL RING.

**Theorem 1.1.13.** *Any ideal $\mathscr{I}$ in $\Pi$ is a principal ideal, more precisely,*
$$\mathscr{I} = \langle \gcd(f : f \in \mathscr{I} \setminus \{0\}) \rangle \tag{1.1.14}$$

**Proof:** For any $f_0, f_1 \in \mathscr{I} \setminus \{0\}$, the Bézout identity (1.1.6) implies that $g_1 = \gcd(f_0, f_1) \in \mathscr{I}$. By divisibility, $f_0, f_1 \in \langle g_1 \rangle$. If $\mathscr{I} \subseteq \langle g_1 \rangle$ we are done, otherwise there exists $f_2 \in (\mathscr{I} \setminus \{0\}) \setminus \langle g_1 \rangle$. Then $g_2 := \gcd(f_2, g_1)$ is a proper divisor of $g_1$ and thus has lower degree than $g_1$. By the same argument as above we know that $g_2 \in \mathscr{I} \setminus \{0\}$ and that $f_0, f_1, f_2 \in \langle g_2 \rangle$. After finitely many repetitions of this process, say $n$ of them, we either have
$$g_n = 1 = \gcd(f : f \in \mathscr{I} \setminus \{0\}) = \gcd(f_1, \ldots, f_n)$$
or
$$(\mathscr{I} \setminus \{0\}) \setminus \langle g_n \rangle = \emptyset$$
and in both cases we can conclude that $\mathscr{I} = \langle g_n \rangle$ and being the generator of the ideal, $g_n$ must be a divisor of all its members. □

Inspecting the proof of Theorem 1.1.13, we see that we proved even more, namely that polynomial ideals have a certain finiteness. This is, of course, a consequence of Hilbert's basissatz that holds even in the multivariate case and a property that is shared by so–called NOETHERIAN RINGS, cf. [Gröbner, 1968, Gröbner, 1970].

**Corollary 1.1.14.** *For any ideal $\mathscr{I} \subset \Pi$ there exists finitely many polynomials $f_1, \ldots, f_n \in \mathscr{I}$ such that $\mathscr{I} = \langle \gcd(f_1, \ldots, f_n) \rangle$.*

Even if the proof of Theorem 1.1.13 is not constructive since we do not know how to choose an element of $(\mathscr{I} \setminus \{0\}) \setminus \langle g_k \rangle$, it is based on an algorithmic concept - or at least this was how we had obtained the Bézout identity. This already justifies the slightly more complicated approach in Algorithm 1.1.2.

Corollary 1.1.14 has another interesting interpretation: finding common zeros of polynomials, in other words, solving a *system* of polynomial equations (in one variable) corresponds

to finding the zeros of the basis element of the ideal. Indeed, common zeros are a property of the ideal,

$$f_1(\xi) = \cdots = f_n(\xi) = 0 \qquad \Leftrightarrow \qquad \left(\sum_{j=1}^{n} p_j f_j\right)(\xi) = 0, \quad p_j \in \Pi,$$

and considering the generator of the (principal) ideal leads to an easier equation, defined by a polynomial of lower degree. This is a concept that will become *very* important in several variables.

### 1.1.3 Zeros of polynomials

A ZERO $\xi$ of a polynomial $f \in \Pi$ is an element $\xi \in \mathbb{K}$ such that $f(\xi) = 0$. This is the point where the field becomes interesting. In finite fields the problem is quite intricate[3] and has applications for example in coding theory, cf. [Cohen et al., 1999, Gathen and Gerhard, 1999], but we are only interested in "real" fields like $\mathbb{Q}$, $\mathbb{R}$ or $\mathbb{C}$ here. These are fields of characteristic zero. Recall that the CHARACTERISTIC of a field $\mathbb{K}$ is the smallest number $n$ such that

$$\underbrace{1 + \cdots + 1}_{n} = 0,$$

or zero if the above never happens. Finite fields have nonzero characteristic.

If a $f(\xi) = 0$ then (1.1.5) with $g = \cdot - \xi$ yields that

$$f(x) = (x - \xi)q(x) + c, \qquad c \in \mathbb{K},$$

and substitution of $\xi$ into this equality implies that $c = 0$. Hence,

$$f(\xi) = 0 \qquad \Leftrightarrow \qquad f = (\cdot - \xi)\, q \qquad \Leftrightarrow \qquad \frac{f}{\cdot - \xi} \in \Pi, \tag{1.1.15}$$

the existence of a zero is equivalent to the existence of a linear factorization. Whether or not a polynomial has zeros or not depends on the underlying field $\mathbb{K}$.

**Example 1.1.15.** The polynomial $x^2 - 2$ has no zeros in $\mathbb{Q}$, but zeros in $\mathbb{R}$ and $\mathbb{C}$, the polynomial $x^2 + 1$ has zeros only in $\mathbb{C}$.

Much of algebraic geometry works over $\mathbb{C}$ and we will also do so, even if it sounds contradictionary: REAL ALGEBRAIC GEOMETRY is significantly more complex, cf. [Basu et al., 2003, Schmüdgen, 2017]. The reason why complex numbers are so popular is the following.

**Theorem 1.1.16.** *The field $\mathbb{C}$ is* ALGEBRAICALLY CLOSED*: for any polynomial $f \in \mathbb{C}[x]$ there exists $\zeta_1, \ldots, \zeta_n$, $n := \deg f$, and $c \in \mathbb{C} \setminus \{0\}$ such that*

$$f = c \prod_{j=1}^{n} (\cdot - \zeta_j). \tag{1.1.16}$$

The proof uses a little bit of FUNCTION THEORY, cf. [Freitag and Busam, 2005, Hille, 1982], namely the fact that polynomials are holomorphic and that holomorphic functions without zeros in $\mathbb{C}$ have to be constant. Hence, as long as $f$ is a non constant polynomials it has at least one zero that can be divided off by (1.1.15), reducing the degree by 1. After finitely many steps one is then left with a constant polynomial and that's it.

What can be done for a single polynomial can be done for a finite number of polynomials as well.

[3] Though in the end it is all combinatorics and could be checked by simply trying all values.

**Definition 1.1.17.** $\xi \in \mathbb{K}$ is called a COMMON ZERO of $f_1, \ldots, f_n \in \Pi$ if

$$0 = f_1(\xi) = \cdots = f_n(\xi). \tag{1.1.17}$$

We write $Z(f_1, \ldots, f_n) \subset \mathbb{K}$ for the set of all common zeros of $f_1, \ldots, f_n$.

**Lemma 1.1.18.** *A point $\xi \in \mathbb{K}$ is a common zero of $f_1, \ldots, f_n$ if and only if it is a zero of* $\gcd(f_1, \ldots, f_n)$.

**Proof:** Any zero of the gcd is a zero of all $f_j = g_j \gcd(f_1, \ldots, f_n)$, and any common zero is a zero of $f_1, f_2$, hence, again by (1.1.6), also of

$$\gcd(f_1, f_2) = p_1 f_1 + p_2 f_2 \qquad \Rightarrow \qquad \gcd(f_1, f_2)(\xi) = p_1(\xi) f_1(\xi) + p_2(\xi) f_2(\xi) = 0.$$

The rest is induction taking into account that

$$\gcd(f_1, \ldots, f_n) = \gcd\big(f_n, \gcd(f_1, \ldots, f_{n-1})\big) = p_n f_n + q_n \gcd(f_1, \ldots, f_{n-1}) = \sum_{j=1}^{n} p_j f_j.$$

□

The ideal theoretic interpretation of the above lemma is even more interesting. The principal ideal $\langle f_1, \ldots, f_n \rangle$ is generated by the BASIS $\{\gcd(f_1, \ldots, f_n)\}$, more precisely,

$$\gcd(f_1, \ldots, f_n) \in \langle f_1, \ldots, f_n \rangle \qquad \text{and} \qquad f_1, \ldots, f_n \in \langle \gcd(f_1, \ldots, f_n) \rangle, \tag{1.1.18}$$

i.e., $\langle f_1, \ldots, f_n \rangle = \langle \gcd(f_1, \ldots, f_n) \rangle$, and all common zeros related to the ideal can be found in this basis element as well. This is a concept that we can and will generalize to the multivariate case and that is the basis for all efficient ideal computations.

The final question is: How can we *compute* zeros of a polynomial? Once we can do that, common zeros are no problem any more as we simply[4] determine the gcd first and then compute its zeros. There is, of course, Newtons method or other analytic zero finding methods, see [Gautschi, 1997, Isaacson and Keller, 1966], but we want to apply a purely algebraic method that reduces the problem to an eigenvalue problem. In that course, we restrict ourselves to the case that the polynomial $f$ has only simple zeros, degree $n+1$ and is monic, i.e.,

$$f(x) = (x - \zeta_0) \cdots (x - \zeta_n), \qquad \zeta_j \in \mathbb{C}, \quad j = 0, \ldots, n. \tag{1.1.19}$$

Being monic is no restriction since the location of the zeros is independent of normalization. Also, we do not care whether $\mathbb{K} = \mathbb{Q}, \mathbb{R}, \mathbb{C}$, the method works within any of these fields and only the eigenvalue problem at the end may lead to complex numbers as also rational or real matrices can have complex eigenvalues.

**Definition 1.1.19.** For $f \in \Pi$, the QUOTIENT SPACE $\Pi / \langle f \rangle$ is the ring defined as $(\Pi)_f$ with the multiplication

$$p \cdot q := \big(pq\big)_f, \tag{1.1.20}$$

It is isomorphic to the vector space

$$\Pi_n := \big\{ g \in \Pi : \deg g \le n \big\}, \qquad n = \deg f - 1. \tag{1.1.21}$$

The difference between $\Pi_n$ and $\Pi / \langle f \rangle$ is that the latter has a well defined multiplication that maps the ring to itself.

[4] This is actually not so simple, especially when done numerically, as can seen in [Corless et al., 2004].

**Example 1.1.20.** For $f = x^2 - x + 2$ we have $\Pi/\langle f \rangle \simeq \Pi_1$, but the product $(x+1)(x-1)$ takes the value $x^2 - 1$ in $\Pi$, which does not belong to $\Pi_1$ any more, while in $\Pi/\langle f \rangle$ we obtain

$$x^2 - 1 - (x^2 - x + 2) = x - 3$$

which is in the quotient space again.

We observe that the multiplication operator $M_f$

$$\Pi/\langle f \rangle \ni p \mapsto M_f p := \big((\cdot)\, p\big)_f, \qquad M_f : \Pi/\langle f \rangle \to \Pi/\langle f \rangle \tag{1.1.22}$$

is a linear operator and therefore its action with respect to an arbitrary basis of $\Pi_n$ can be expressed by means of a matrix.

**Definition 1.1.21.** Let $B = \{b_0, \ldots, b_n\}$ be any basis of $\Pi_n$. The matrix

$$M_B(f) = \Big(m_{jk}^B(f) : j,k = 0,\ldots,n\Big) \in \mathbb{K}^{n+1\times n+1}$$

defined by

$$M_f b_j = \sum_{k=0}^{n} m_{jk}^B(f)\, b_k, \qquad j = 0,\ldots,n,$$

is called the COMPANION MATRIX of $f$ with respect to the basis $B$.

Clearly, the companion matrix depends on the choice of the basis $B$ as can be seen in the following example.

**Example 1.1.22.** Let $f = x^{n+1} + f_{n-1}x^{n-1} + \cdots + f_0 \in \mathbb{K}[x]$, where $\mathbb{K} = \mathbb{Q}, \mathbb{R}, \mathbb{C}$.

1. If $B = \{1, x, \ldots, x^n\}$, then
$$M_B(f) = \begin{pmatrix} 0 & & & -f_0 \\ 1 & & & -f_1 \\ & \ddots & & \vdots \\ & & 1 & -f_n \end{pmatrix}.$$
This is the so–called FROBENIUS COMPANION MATRIX.

2. If $b_0, \ldots, b_n$ are a monic orthogonal polynomials with respect to any inner product $(;\cdot)$ such that $(f, \Pi_n) = 0$, the companion matrix takes the form
$$M_B(f) = \begin{pmatrix} -\beta_1 & \gamma_2 & & & \\ 1 & -\beta_2 & \gamma_3 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -\beta_n & \gamma_{n+1} \\ & & & 1 & -\beta_{n+1} \end{pmatrix},$$
where the $\beta_j$ and $\gamma_j$ are the coefficients in the three term recurrence. This is used for the computation on Gaussian quadrature nodes, cf. [Gautschi, 1997, Sauer, 2019].

3. For $\xi_0, \ldots, \xi_n \in \mathbb{K}$, using the basis $b_j = (\cdot - \xi_0)\cdots\big(\cdot - \xi_{j-1}\big)$, $j = 0,\ldots,n$, we get the companion matrix
$$M_B(f) = \begin{pmatrix} \xi_0 & & & -\frac{[\xi_0]f}{[\xi_0,\ldots,\xi_{n+1}]f} \\ 1 & \xi_1 & & -\frac{[\xi_0,\xi_1]f}{[\xi_0,\ldots,\xi_{n+1}]f} \\ & \ddots & \ddots & \vdots \\ & & 1 & \xi_n - \frac{[\xi_0,\ldots,\xi_n]f}{[\xi_0,\ldots,\xi_{n+1}]f} \end{pmatrix}$$

based on divided differences of $f$, first considered in [Calvetti et al., 2003]. If, by accident, $\xi_j = \zeta_j$, then this matrix just consists of the diagonal and the subdiagonal.

More on these matrices and what can be done with them can be found in [Sauer, 2018a].

The main observation of the multiplication operator modulo $f$ is now as follows. We define

$$\ell_j = \prod_{k \neq j} (\cdot - \zeta_k) = \frac{f}{\cdot - \zeta_j}, \qquad j = 0, \dots, n$$

and only note that, in $\Pi / \langle f \rangle$,

$$0 \equiv f = (\cdot - \zeta_j)\, \ell_j = M_f \ell_j - \zeta_j \ell_j, \qquad j = 0, \dots, n,$$

hence

$$M_f \ell_j = \zeta_j \ell_j, \qquad j = 0, \dots, n. \tag{1.1.23}$$

This already proves the following theorem.

**Theorem 1.1.23.** *If $f$ is of the form* (1.1.19)*, its zeros are exactly the eigenvalues of the multiplication operator and therefore of any companion matrix.*

**Remark 1.1.24.** Even if the eigenvalues are the same, their numerical conditioning can be different for different companion matrices and it can make sense to vary the basis.

There remains the question what to do if if $f$ has multiple zeros. But this is easy since $\gcd(f, f')$ has the same zeros as $f$, however as simple zeros.

## 1.2 Numerical applications

We next review some numerical applications that involve polynomials and how they look in one variable. The rest of the lecture will be used to generalize these applications to several variables.

### 1.2.1 Polynomial Interpolation

Interpolation is one of the most classical numerical problem, even the name already dates back to Wallis in 1655, see [Bauschinger, 1900]. The interpolation problem is as follows: given SITES $x_j \in \mathbb{R}$ and values $y_j \in \mathbb{R}$, $j = 0, \dots, n$, find a function $f$ such that

$$f(x_j) = y_j, \qquad j = 0, \dots, n. \tag{1.2.1}$$

In this generality, the problem has infinitely many solution, even if we require the function to be continuous or to have some order of differentiability. Proper polynomials, on the other hand, solve the problem.

**Theorem 1.2.1.** *If $x_j \in \mathbb{R}$, $j = 0, \dots, n$, are pairwise disjoint sites, then there exists, for any $y_j \in \mathbb{R}$, $j = 0, \dots, n$, a **unique** polynomial $f \in \Pi_n$ such that* (1.2.1) *is satisfied.*

**Proof:** That there exists a solution can be seen from the explicit formula

$$f(x) = \sum_{j=0}^{n} y_j \prod_{k \neq j} \frac{x - x_k}{x_j - x_k}, \tag{1.2.2}$$

and uniqueness follows since the difference of two interpolants $f, g$ vanishes at all $x_j$, hence

$$f - g = q\,(\cdot - x_0) \cdots (\cdot - x_n), \qquad q \in \Pi,$$

yielding $q = 0$ and $f = g$ as otherwise the polynomial on the left hand side has degree $\leq n$, the one the right hand side degree $\geq n+1$. □

**Remark 1.2.2.** Polynomial interpolation problems are usually formulated over $\mathbb{R}$, but in fact the field is irrelevant and the argument of the proof of Theorem 1.2.1 works in any field, even in finite ones.

A particular role is played be the polynomial

$$\omega = \prod_{j=0}^{n} (\cdot - x_j) \in \Pi_{n+1} \tag{1.2.3}$$

that vanishes at all the sites. It allows us to rewrite (1.2.2) as

$$f = \sum_{j=0}^{n} y_j \frac{\omega}{(\cdot - x_j)\,\omega'(x_j)},$$

a classical formula used for example by Gauss in [Gauss, 1816], see [Sauer, 2019]. To study the action of interpolation on polynomials, we make the following definition.

**Definition 1.2.3.** Given $\mathscr{X} = \{x_0, \ldots, x_n\}$, the INTERPOLATION OPERATOR $L_{\mathscr{X}} : \Pi \to \Pi_n$ is defined by $L_{\mathscr{X}} f(\mathscr{X}) = f(\mathscr{X})$, $f \in \Pi$.

If we take any $f \in \Pi$ and write it in the form

$$f = q\,\omega + r, \qquad r = (f)_\omega,$$

a simple substitution yields that $f(x_j) = r(x_j)$, $j = 0, \ldots, n$, hence

$$L_{\mathscr{X}} = (\cdot)_\omega \tag{1.2.4}$$

is simply a REMAINDER of division. Moreover, $L_{\mathscr{X}}$ is a PROJECTION OPERATOR, i.e.,

$$L_{\mathscr{X}} \left( L_{\mathscr{X}} f \right) = L_{\mathscr{X}} f,$$

and satisfies

$$\ker L_{\mathscr{X}} = \left\{ f \in \Pi : f(\mathscr{X}) = 0 \right\} = \langle \omega \rangle,$$

which is why polynomial interpolation is called an IDEAL PROJECTOR.

**Remark 1.2.4.** There are two interesting aspects of the simple identity (1.2.4):

1. If we interpret it in terms of ideals, the interpolant is simply the represented modulo ideal in *any* quotient space $\Pi/\mathscr{I}$, regardless of how we choose it[5]; this will become relevant in several variables when there is no canonical representer like $\Pi_n$ and things depend on geometry and not only on counting.

[5] Keep in mind that the space is only defined up to adding multiples of $\omega$ to its basis elements. The canonical choice $\Pi_n$ corresponds to using *zero* multiples.

2. The formula also admits multiple points since

$$\omega = (\cdot - x_0)^{\mu_0} \cdots (\cdot - x_n)^{\mu_n}, \qquad \mu_j \in \mathbb{N},\ j = 1, \ldots, n,$$

is also defined if the MULTIPLICITY $\mu_k$ of some site is $> 1$. The remainder, hence the interpolant, is now of degree $\mu_0 + \cdots + \mu_n - 1$ and since

$$\left(\frac{d^k}{dx^k}\omega\right)(x_j) = 0 \qquad k = 0, \ldots, \mu_j - 1, \quad j = 0, \ldots, n,$$

it follows for any $f \in \Pi$, $k = 0, \ldots, \mu_j - 1$ and $j = 0, \ldots, n$ that

$$\left(\frac{d^k}{dx^k} r\right)(x_j) = \left(\frac{d^k}{dx^k} f\right)(x_j) - \sum_{\ell=0}^{k} \binom{k}{\ell} \underbrace{\left(\frac{d^k}{dx^k}\omega\right)(x_j)}_{=0} \left(\frac{d^k}{dx^k} q\right)(x_j) = \left(\frac{d^k}{dx^k} f\right)(x_j),$$

hence the remainder performs HERMITE INTERPOLATION, interpolating $\mu_j$ *consecutive* derivatives at $x_j$.

One can use interpolation to bound the degrees in the Bézout coefficients. To that end, let, for simplicity, $f, g$ be polynomials with simple zeros, for multiple zeros, Hermite interpolation would have to be incorporated. Then we can show the following, cf. [DeVilliers et al., 2000].

**Proposition 1.2.5.** *Let $f, g \in \Pi$ be two polynomials of degree $m, n$, respectively, with simple zeros. Then there exist $p \in \Pi_{n-1}$ and $q \in \Pi_{m-1}$ such that* $\gcd(f, g) = pf + qg$.

**Proof:** Write the polynomials in factorized form

$$f(x) = f_m \prod_{j=1}^{m} \left(x - \zeta_j\right), \qquad g(x) = g_n \prod_{j=1}^{n} \left(x - \zeta'_j\right), \qquad \gcd(f, g)(x) = \prod_{j=1}^{k} \left(x - \zeta_j\right),$$

hence $\zeta_j = \zeta'_j$, $j = 1, \ldots, k$. Now, we denote by $\tilde{p} \in \Pi_{n-k-1}$ and $\tilde{q} \in \Pi_{m-k-1}$ the unique solutions of the interpolation problems

$$p(\xi'_j) = \frac{1}{f(\zeta'_j)}, \quad j = k+1, \ldots, n, \qquad q(\xi_j) = \frac{1}{g(\zeta_j)}, \quad j = k+1, \ldots, m,$$

then

$$p\frac{f}{\gcd(f, g)} + q\frac{g}{\gcd(f, g)} \in \Pi_{n+m-2k-1}$$

takes the value 1 at the the $n + m - 2k$ point $\zeta_j$, $j = k+1, \ldots, m$ and $\zeta'_j$, $j = k+1, \ldots, n$, hence, by uniqueness of interpolation must be the constant polynomial 1. Then,

$$\gcd(f, g) = \gcd(f, g)\left(p\frac{f}{\gcd(f, g)} + q\frac{g}{\gcd(f, g)}\right) = pf + qg,$$

which proves the claim. □

**Corollary 1.2.6.** *The degrees of $p$ and $q$ in Proposition 1.2.5 can even be chosen as $n - 1 - \deg\gcd(f, g)$ and $m - 1 - \deg\gcd(f, g)$, respectively.*

A different approach to interpolation is by pure linear algebra and uses only the vector space properties of polynomials. If $\Phi = \{\phi_0, \dots, \phi_n\}$ is a basis of a finite dimensional space of functions, at least defined on $\mathscr{X}$, then the interpolant to $\mathscr{X}$ from this space can be written as

$$\phi = \sum_{k=0}^{n} a_k \phi_k, \qquad a_k \in \mathbb{K},$$

and the interpolation problem takes the form

$$y_j = \phi(x_j) = \sum_{k=0}^{n} a_k \phi_k(x_j) = e_j^T \begin{pmatrix} \phi_0(x_0) & \dots & \phi_n(x_0) \\ \vdots & \ddots & \vdots \\ \phi_0(x_n) & \dots & \phi_n(x_n) \end{pmatrix} \begin{pmatrix} a_0 \\ \vdots \\ a_n \end{pmatrix}, \qquad j = 0, \dots, n,$$

or, more compactly

$$y = \big(\phi_k(x_j) : j, k = 0, \dots, n\big)\, a, \tag{1.2.5}$$

which corresponds to solving a linear system[6] and (unique) solvability of the interpolation problem can be decided entirely by linear algebra. The matrix in this system even has a special name.

**Definition 1.2.7.** Given a finite set $\mathscr{X} \subset \mathbb{R}$ and a finite set $\Phi$ of functions $\mathscr{X} \to \mathbb{R}$, the associated COLLOCATION MATRIX is defined as

$$V(\mathscr{X}, \Phi) := \left(\phi(x) : \begin{array}{c} x \in \mathscr{X} \\ \phi \in \Phi \end{array}\right). \tag{1.2.6}$$

In the case that $\Phi = \{1, x, \dots, x^n\}$ the matrix is called VANDERMONDE MATRIX, denoted by

$$V_n(\mathscr{X}) := V\big(\mathscr{X}, \{1, \dots, x^n\}\big). \tag{1.2.7}$$

**Remark 1.2.8.** The terminology "collocation matrix" and "Vandermonde matrix" is not consistent in the literature, often the matrix from (1.2.6) is called Vandermonde matrix even for nonpolynomial systems or for a different basis of $\Pi_n$.

Note that if $\Phi, \Phi'$ are two bases of the same space, then there exists a nonsingular matrix $A$ such that $\Phi' = A\Phi$ and hence

$$V(\mathscr{X}, \Phi') = V(\mathscr{X}, \Phi)\, A,$$

so that collocation matrices are relatively invariant under changes of basis and essentially depend on the space.

Linear algebra tells us that the interpolation problem has a unique solution if and only if $V(\mathscr{X}, \Phi)$ is invertible which first implies that the matrix is a square one, i.e., $\#\mathscr{X} = \#\Phi$, and that $\det V(\mathscr{X}, \Phi) \neq 0$. Now, we can derive the unique solvability of the polynomial interpolation problem by different means.

**Theorem 1.2.9.** *There exists $c \neq 0$ such that*

$$\det V_n(\mathscr{X}) = c \prod_{j \neq k} \big(x_j - x_k\big), \qquad \mathscr{X} = \{x_0, \dots, x_k\}. \tag{1.2.8}$$

[6] Surprisingly, this fact is rediscovered regularly, especially in the context of multivariate interpolation. That does not increase its novelty, unfortunately.

**Proof:** We first note that $f(x_0,\dots,x_n) := \det V_n(\mathscr{X})$ is a polynomial of (total) degree[7] $\frac{1}{2}n(n+1)$. This is a simple induction based on expanding the determinant with respect to the last column:

$$\begin{aligned} f(x_0,\dots,x_n) &= \begin{vmatrix} 1 & x_0 & \dots & x_0^n \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \dots & x_n^n \end{vmatrix} \\ &= (-1)^n \left( x_0^n \begin{vmatrix} 1 & x_1 & \dots & x_1^{n-1}n \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \dots & x_n^{n-1}n \end{vmatrix} + \dots + (-1)^n x_n^n \begin{vmatrix} 1 & x_0 & \dots & x_0^{n-1}n \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n-1} & \dots & x_{n-1}^{n-1} \end{vmatrix} \right) \\ &= \sum_{j=0}^{n} (-1)^{n+j} x_j^n \det V_{n-1}\left(\mathscr{X} \setminus \{x_j\}\right). \end{aligned}$$

Moreover, $f$ is a polynomial of degree $n$ in each individual variable $x_j$ that vanishes in $x_k$, $k \neq j$, as then the matrix has to identical rows. Hence, for any $j = 0,\dots,n$,

$$f(x_0,\dots,x_n) = \left( \prod_{k\neq j} (x_j - x_k) \right) g\left(x_0,\dots,x_{j-1},x_{j+1},\dots,x_n\right),$$

hence any $(x_j - x_k)$, $j \neq k$, divides $f$ and therefore there exists a polynomial $g$ such that

$$f(x_0,\dots,x_n) = g(x_0,\dots,x_n) \prod_{j\neq k} (x_j - x_k),$$

and since the product on the right has $\frac{1}{2}n(n+1)$ factors, hence degree $\frac{1}{2}n(n+1)$, $g$ must be a constant polynomial. That the constant is nonzero follows from Theorem 1.2.1. □

**Remark 1.2.10.** Historically, polynomial interpolation in one variable is a more than classical issue, investigate for example by Newton in his *Principia* where the NEWTON FORMULA for the interpolant has been derived[8]. A nice summary has been given in [Bauschinger, 1900], later translated into French in [Andoyer, 1906] by Andoyer who missed the difficulties of the multivariate case, cf. [Gasca and Sauer, 2000b]. Multivariate interpolation, on the other hand, is a fairly recent issue except a few results by Jacobi [Jacobi, 1835] and Kronecker [Kronecker, 1866] in the 19th century.

### 1.2.2 Signal processing and generating functions

SIGNAL PROCESSING, more precisely, digital signal processing in the sense of [Hamming, 1989] is concerned with the action of filters on discrete signals.

**Definition 1.2.11** (Spaces & filters)**.**

1. By $\ell(\mathbb{Z})$ we denote the space of all INFINITE SEQUENCES, i.e., all functions $\mathbb{Z} \to \mathbb{R}$. Moreover, $\ell_p(\mathbb{Z})$, $0 \le p \le \infty$, stands for all sequences such that the "norm"[9]

$$\|c\|_p := \left( \sum_{k\in\mathbb{Z}} |c(k)|^p \right)^{1/p}, \qquad \|c\|_\infty := \sup_{k\in\mathbb{Z}} |c(k)|, \qquad \|c\|_0 := \#\{k : c(k) \neq 0\},$$

[7] More on degrees for multivariate polynomials later, now only so much: the TOTAL DEGREE of a monomial $x_0^{\alpha_0}\cdots x_n^{\alpha_n}$ is $\alpha_0 + \dots + \alpha_n$ and the total degree of a polynomial the maximal total degree of its monomial components.

[8] It is by no means to expect that something that carries someones name has really been invented by this person, so it is a worthwhile remark

[9] It is only a NORM for $1 \le p \le \infty$, otherwise a so called QUASI NORM.

is finite.

2. For $c, d \in \ell_1(\mathbb{Z})$ we define the CONVOLUTION

$$c * d := \sum_{k\in\mathbb{Z}} c(\cdot - k)\, d(k).$$

3. A FILTER $F : \ell(\mathbb{Z}) \to \ell(\mathbb{Z})$ is a convolution operator

$$Fc = f * c, \qquad c \in \ell(\mathbb{Z}),$$

with an IMPULSE RESPONSE $f \in \ell_0(\mathbb{Z})$.

The name *impulse response* is easily explained. If we use the PEAK SEQUENCE $\delta \in \ell(\mathbb{Z})$, $\delta(k) = \delta_{k0}$, $k \in \mathbb{Z}$, which is the identity in the CONVOLUTION ALGEBRA[10], then $f = F\delta$, hence, in a slightly ugly notation, $F = (F\delta)*$.

Filters and convolutions can be put into an algebraic framework by turning sequences into Laurent polynomials.

**Definition 1.2.12.** Let $c \in \ell_0(\mathbb{Z})$. The $z$–TRANSFORM $c^\flat$ and the SYMBOL or GENERATING FUNCTION $c^\sharp$ are defined as

$$c^\flat(z) = \sum_{k\in\mathbb{Z}} c(k)\, z^{-k}, \qquad c^\sharp(z) = \sum_{k\in\mathbb{Z}} c(k)\, z^{k}, \qquad z \in \mathbb{C}_\times := \mathbb{C} \setminus \{0\}. \tag{1.2.9}$$

**Remark 1.2.13.** In principle, all the theory can be made in terms of $z$–transforms or symbols as, due to

$$c^\flat = c^\sharp\left((\cdot)^{-1}\right) \qquad \text{and} \qquad c^\sharp = c^\flat\left((\cdot)^{-1}\right)$$

they are almost completely equivalent except a few minor issues like the inverse $z$–transform, cf. [Föllinger, 2000]. The $z$–transform is more common in signal processing while symbols are more frequently used in mathematics. And generating functions are even meaningful for infinitely supported sequences, but have to be handled with a bit of care.

The important of $z$–transforms and symbols becomes evident from the simple observation that

$$\begin{aligned}(c * d)^\flat(z) &= \sum_{j\in\mathbb{Z}} (c*d)_k z^{-j} = \sum_{k\in\mathbb{Z}}\sum_{j\in\mathbb{Z}} c(j-k)\, d(k) z^{k-j} z^{-k} = \sum_{k\in\mathbb{Z}} d(k)\, z^{-k} \sum_{j\in\mathbb{Z}} c(j-k)\, z^{k-j} \\ &= \sum_{k\in\mathbb{Z}} d(k)\, z^{-k} \sum_{j\in\mathbb{Z}} c(j)\, z^{-j},\end{aligned}$$

hence,

$$(c*d)^\flat = c^\flat d^\flat \qquad \text{and} \qquad (c*d)^\sharp = (c*d)^\flat\left((\cdot)^{-1}\right) c^\flat\left((\cdot)^{-1}\right) d^\flat\left((\cdot)^{-1}\right) = c^\sharp d^\sharp. \tag{1.2.10}$$

In other words, the transforms modify the rater complicated operation of convolution into a simple product of Laurent polynomials, an operation well compatible with the ring structure of $\Lambda$.

This algebraization becomes particularly useful when passing to filterbanks which combine several filters to decompose a signal into several bands. To maintain the ratio between the amount of data and the information contained in it, filterbanks use decimations.

[10] (Finitely supported) sequences with componentwise addition and convolution as a multiplication.

**Definition 1.2.14** (Down- & upsampling)**.** Given a dilation factor $m \in \mathbb{N}$, the DOWNSAMPLING OPERATOR $\downarrow_m$ is defined as

$$\downarrow_m : \ell(\mathbb{Z}) \to \ell(\mathbb{Z}), \qquad c \mapsto c(m\cdot) \tag{1.2.11}$$

and the UPSAMPLING OPERATOR as

$$\uparrow_m : \ell(\mathbb{Z}) \to \ell(\mathbb{Z}), \qquad c(k) = \begin{cases} c(k/m), & k \in m\mathbb{Z}, \\ 0, & k \notin m\mathbb{Z}, \end{cases} \quad k \in \mathbb{Z}. \tag{1.2.12}$$

Moreover, we use the notation $\mathbb{Z}_m = \mathbb{Z}/m\mathbb{Z} \simeq \{0, \dots, m-1\}$.

**Exercise 1.2.1** Show that $\downarrow_m \uparrow_m = I \neq \uparrow_m \downarrow_m$ and that

$$I = \sum_{j \in \mathbb{Z}_m} \tau^j \uparrow_m \downarrow \tau^{-j}. \tag{1.2.13}$$

♢

Up- and downsampling can be computed in terms of $z$–transform, upsampling simply by noting that

$$(\uparrow_m c)^\flat(z) = \sum_{k \in \mathbb{Z}} c(k)\, z^{-mk} = c^\flat\left(z^m\right), \tag{1.2.14}$$

while downsampling is based on the FOURIER IDENTITY

$$\frac{1}{m} \sum_{k \in \mathbb{Z}_m} e^{-2\pi i j k/m} = \begin{cases} 1, & j \in m\mathbb{Z}, \\ 0, & j \notin m\mathbb{Z}, \end{cases} \qquad j \in \mathbb{Z}, \tag{1.2.15}$$

which gives that

$$\begin{aligned} (\downarrow_m c)^\flat(z^m) &= \sum_{j \in \mathbb{Z}} c(mj) z^{-mj} = \sum_{j \in \mathbb{Z}} c(j) z^{-j} \frac{1}{m} \sum_{k \in \mathbb{Z}_m} e^{-2\pi i j k/m} \\ &= \frac{1}{m} \sum_{k \in \mathbb{Z}_m} \sum_{j \in \mathbb{Z}} c(j) \left(e^{-2\pi i k/m} z\right)^{-j} = \frac{1}{m} \sum_{k \in \mathbb{Z}_m} c^\flat\left(e^{2\pi i k/m} z\right). \end{aligned} \tag{1.2.16}$$

Note that the UNIT ROOTS $e^{2\pi i k/m}$, $k \in \mathbb{Z}_m$, can be seen as generalized signs; in the case of $m = 2$ they are indeed $\pm 1$.

**Exercise 1.2.2** Prove (1.2.15) or at least find a proof of it in the literature. ♢

Now we can define the concept of a *univariate* filterbank, cf. [Vetterli and Kovačević, 1995].

**Definition 1.2.15** (Filterbank)**.** A FILTERBANK consists of $n$ ANALYSIS FILTERS $F_j$ with impulse response $f_j$, $j \in \mathbb{Z}_n$, and $n$ SYNTHESIS FILTERS $G_j$ with impulse response $g_j$, $j \in \mathbb{Z}_n$. It computes the SUBBAND DECOMPOSITION

$$\ell(\mathbb{Z}) \ni c \mapsto Fc := \left[\downarrow_m F_j c : j \in \mathbb{Z}_n\right] = \begin{pmatrix} \downarrow_m (f_0 * c) \\ \vdots \\ \downarrow_m (f_{n-1} * c) \end{pmatrix} \in \ell(\mathbb{Z})^n$$

and the SUBBAND RECONSTRUCTION

$$\ell(\mathbb{Z})^n \ni c = \begin{pmatrix} c_0 \\ \vdots \\ c_{n-1} \end{pmatrix} \mapsto Gc = \sum_{j \in \mathbb{Z}_n} G_j \uparrow_m c_j = \sum_{j \in \mathbb{Z}_n} g_j * \left(\uparrow_m c_j\right).$$

The filterbank is said to

1. provide PERFECT RECONSTRUCTION if $GF = I$,
2. be CRITICALLY SAMPLED if $m = n$.

Filterbanks without perfect reconstruction loose information and thus make no sense, and since the decimation $\downarrow_m$ in the analysis part just compensates the $n$ components of the vectorized subband data, critically sampled filterbanks are by far the most popular ones. The analysis filterbank can be depicted as

$$\begin{array}{ccccccc} & \nearrow & F_0 & \to & \downarrow_m & \to & c_0 \\ c & & \vdots & & \vdots & & \vdots \\ & \searrow & F_{n-1} & \to & \downarrow_m & \to & c_{n-1} \end{array} \tag{1.2.17}$$

and the synthesis filterbank takes the form

$$\begin{array}{ccccccc} c_0 & \to & \uparrow_m & \to & G_0 & \searrow & \\ \vdots & & \vdots & & \vdots & & \oplus \to c \\ c_{n-1} & \to & \uparrow_m & \to & G_{n-1} & \nearrow & \end{array} \tag{1.2.18}$$

We can now describe the action of the filterbank by means of our algebraic tools. Beginning with the synthesis part, we note that, by (1.2.16) and (1.2.10)

$$c_j^\flat(z^m) = \left(\downarrow_m (f_j c)\right)^\flat (z^m) = \frac{1}{m} \sum_{k\in\mathbb{Z}_m} f_j^\flat\left(e^{-2\pi i k/m} z\right) c^\flat\left(e^{-2\pi i k/m} z\right), \qquad j \in \mathbb{Z}_n,$$

which can be vectorized as

$$\begin{pmatrix} c_0^\flat(z^m) \\ \vdots \\ c_{n-1}^\flat(z^m) \end{pmatrix} = \frac{1}{m} \left[ f_j^\flat\left(e^{-2\pi i k/m} z\right) : \begin{array}{c} j \in \mathbb{Z}_n \\ k \in \mathbb{Z}_m \end{array} \right] \begin{pmatrix} c^\flat(z) \\ c^\flat\left(e^{-2\pi i k/m} z\right) \\ \vdots \\ c^\flat\left(e^{-2\pi i (m-1)/m} z\right) \end{pmatrix} \tag{1.2.19}$$

**Definition 1.2.16** (Modulation & polyphase)**.** The matrix

$$\boldsymbol{F}(z) = \frac{1}{m} \left( f_j^\flat\left(e^{-2\pi i k/m} z\right) : \begin{array}{c} j \in \mathbb{Z}_n \\ k \in \mathbb{Z}_m \end{array} \right) \in \Lambda^{n\times m} \tag{1.2.20}$$

is called the ANALYSIS MODULATION MATRIX of the filterbank, the vector

$$\boldsymbol{c}_p^\flat(z) := \left[ c^\flat\left(e^{-2\pi i k/m} z\right) : k \in \mathbb{Z}_m \right] \in \Lambda^m \tag{1.2.21}$$

is called the POLYPHASE VECTOR of the signal $c$. The SUBBAND DECOMPOSITION then takes the form

$$\begin{pmatrix} c_0^\flat(z^m) \\ \vdots \\ c_{n-1}^\flat(z^m) \end{pmatrix} = \boldsymbol{F}(z)\, \boldsymbol{c}_p^\flat(z). \tag{1.2.22}$$

Synthesis is simpler. By (1.2.10) and (1.2.14) we have that

$$\begin{aligned} \tilde{c}^\flat(z) &:= \sum_{k\in\mathbb{Z}_n} \left(g_k * (\uparrow_m c_k)\right)^\flat (z) = \sum_{k\in\mathbb{Z}_n} g_k^\flat(z)\, (\uparrow_m c_k)^\flat (z) = \sum_{k\in\mathbb{Z}_n} g_k^\flat(z)\, c_k(z^m) \\ &= \left(g_0^\flat(z), \ldots, g_{n-1}^\flat(z)\right) \begin{pmatrix} c_0^\flat(z^m) \\ \vdots \\ c_{n-1}^\flat(z^m) \end{pmatrix}. \end{aligned}$$

Building the polyphase vector of the output $\tilde{c}$ and keeping in mind that $\left(e^{-2\pi i j/m}\right)^m = e^{-2\pi i j} = 1$, $j \in \mathbb{Z}_m$, we thus get that

$$\tilde{\boldsymbol{c}}_p^\flat(z) = \left[ g_k^\flat\left(e^{-2\pi i j/m}\right) : \begin{array}{c} j \in \mathbb{Z}_m \\ k \in \mathbb{Z}_n \end{array} \right] \begin{pmatrix} c_0^\flat(z^m) \\ \vdots \\ c_{n-1}^\flat(z^m) \end{pmatrix} =: \boldsymbol{G}^T(z) \begin{pmatrix} c_0^\flat(z^m) \\ \vdots \\ c_{n-1}^\flat(z^m) \end{pmatrix}. \tag{1.2.23}$$

Therefore, decomposition and immediate reconstruction yield

$$\tilde{\boldsymbol{c}}_p^\flat(z) = \boldsymbol{G}(z)^T \boldsymbol{F}(z)\, \boldsymbol{c}_p^\flat(z)$$

and we get the fundamental theorem for the modulation matrices of a perfect reconstruction filterbank.

**Theorem 1.2.17.** *The filterbank provides perfect reconstruction if and only if* $\boldsymbol{G}^T(z)\boldsymbol{F}(z) = I$.

**Remark 1.2.18.** Eventually, the algebraic approach to filterbanks consists of encoding the synthesis and the analysis part of the filterbank as matrices whose components are $z$–transforms, hence Laurent polynomials. Since

$$\big(a_{jk}(z) : j,k\big) = \left( \sum_{\ell\in\mathbb{Z}} a_{jk}(\ell)\, z^\ell : j,k \right) = \sum_{\ell\in\mathbb{Z}} \big(a_{jk}(\ell) : j,k\big)\, z^\ell = \sum_{\ell\in\mathbb{Z}} A_\ell z^\ell,$$

such a matrix of Laurent polynomials can also be seen as a matrix valued Laurent polynomial or a Laurent polynomial with matrix coefficients and thus as the $z$-transform or symbol of a matrix valued sequence.

In a critically sampled filterbank we thus get perfect reconstruction iff, given $\boldsymbol{F}(z)$, we set[11] $\boldsymbol{G}(z) = \boldsymbol{F}^{-1}(z)$. In other words, in this case the requirement of perfect reconstruction implies that the analysis part completely determines the synthesis part and vice versa. Mathematically, this brings us to the question of when a matrix over a ring is invertible in the ring. This is surprisingly simple.

**Proposition 1.2.19.** *A square matrix* $A \in R^{n\times n}$ *has an inverse in* $R^{n\times n}$ *if and only if* $\det A \in R^\times$ *is a unit in* $R$.

**Proof:** If $A$ has an inverse, then

$$\det A \det A^{-1} = \det I = 1,$$

hence $\det A^{-1} = (\det A)^{-1}$ in $R$. The converse follows from Cramer's rule which says that

$$\left(A^{-1}\right)_{jk} = (\det A)^{-1} \det A_{jk}, \qquad j,k = 1,\dots,n,$$

where $A_{jk} \in R^{(n-1)\times(n-1)}$ stands for the matrix where the $j$th row and the $k$th column of $A$ are deleted. □

**Remark 1.2.20.** Theorem 1.2.17 has a different meaning depending on whether we want to choose $\boldsymbol{F}, \boldsymbol{G}$ as polynomials or as Laurent polynomials, which plays a role if one is interested in a causal filter. In the first case, an analysis filterbank can be completed if $\det \boldsymbol{F}(z) \in \mathbb{C}_\times$, in the second case if $\det \boldsymbol{F}(z) = c\, z^k$, $c \in \mathbb{C}_\times$, $k \in \mathbb{Z}$.

[11] We can first do it pointwise but then have to ensure that the resulting function is a matrix of Laurent polynomials again - this is the nontrivial fun part of it.

Hence, the question of constructing reasonable filterbanks boils down to construction a reasonable $\boldsymbol{F}$ or $\boldsymbol{G}$ as the other one follows directly.

**Example 1.2.21.** We consider $n = m = 2$ and want to build

$$\boldsymbol{G}(z) = \begin{pmatrix} g_0^\flat(z) & g_1^\flat(z) \\ g_0^\flat(-z) & g_1^\flat(-z) \end{pmatrix}$$

with $\det \boldsymbol{G}(z) = 1$. This is impossible if there exists $z^*$ such that $g_0^\flat(z^*) = g_0^\flat(-z^*) = 0$ as then $z - z^*$ divides $\det \boldsymbol{G}(z)$. If, on the other hand $\gcd\left(g_0^\flat, g_0^\flat(-\cdot)\right) = 1$, then there exist $p, q$ such that

$$p(z)\, g_0^\flat(z) + q(z)\, g_0^\flat(-z) = 1, \tag{1.2.24}$$

hence, replacing $z$ by $-z$ in (1.2.24), also

$$p(-z)\, g_0^\flat(-z) + q(-z)\, g_0^\flat(z) = 1$$

and therefore

$$\frac{p(z) + q(-z)}{2} g_0^\flat(z) + \frac{p(-z) + q(z)}{2} g_0^\flat(-z) = 1$$

and setting $g_1^\flat(z) = \frac{p(z)+q(-z)}{2}$ we obtain that $\det \boldsymbol{G}(z) = 1$. This can be summarized as follows: *A filter with impulse response $g_0$ can be completed to a perfect reconstruction filter bank if and only if* $\gcd\left(g_0^\flat, g_0^\flat(-\cdot)\right) = 1$.

### 1.2.3 Subdivision, differences and wavelets

Subdivision, is an iterative way to generate functions from discrete data by means of stationary operators, i.e. operators that do the same thing regardless of where it happens. Actually, these operators are even convolutions. The fundamental monograph on subdivision is still [Cavaretta et al., 1991], for a different, more geometric approach see [Peters and Reif, 2008, Warren, 2001].

We start with a discrete sequence $c \in \ell(\mathbb{Z})$ that we want to extend to a function $c' : \frac{1}{2}\mathbb{Z} \to \mathbb{R}$ by means of local and position independent rules. To that end, we choose two finitely supported filters $a_0$ and $a_1$ and define

$$c'\left(\frac{2j+\epsilon}{2}\right) := \sum_{k\in\mathbb{Z}} a_\epsilon(k)\, c(j-k) = (a_\epsilon * c)(j), \qquad j \in \mathbb{Z}, \qquad \epsilon \in \{0,1\}, \tag{1.2.25}$$

so that $a_1$ is an INSERTION RULE that decides what happens at the new half-integer points and $a_0$ is a REPLACEMENT RULE determining how the integer values are handled – $a_0 = \delta$ would just leave them unchanged, such a subdivision operator is called INTERPOLATORY. To be able to iterate the scheme, we renormalize $c'$ to a sequence $\mathbb{Z} \to \mathbb{R}$, merge the two sequences into one and define the following object.

**Definition 1.2.22** (Subdivision operator)**.** The SUBDIVISION OPERATOR with respect to the MASK $a \in \ell_0(\mathbb{Z})$ is defined as

$$S_a c := \sum_{k\in\mathbb{Z}} a(\cdot - 2k)\, c(k). \tag{1.2.26}$$

Indeed, if we set $a(2\cdot+\epsilon) := a_\epsilon$, $\epsilon \in \{0,1\}$, which uniquely connects $a$ and $a_0, a_1$, we see that

$$S_a c(2\cdot+\epsilon) = \sum_{k\in\mathbb{Z}} a(2\cdot+\epsilon-2k)\, c(k) = \sum_{k\in\mathbb{Z}} a_\epsilon(\cdot-k)\, c(k) = c'\left(j+\frac{\epsilon}{2}\right),$$

and Definition 1.2.22 makes sense as $S_a c = c'(2\cdot)$, that is, as a renormalization of the above geometrically intuitive process.

**Remark 1.2.23.** In terms of signal processing, the subdivision operator is actually a déjà-vu as soon as we write it as a GENERALIZED CONVOLUTION

$$c *_m d := \sum_{j\in\mathbb{Z}} c(\cdot - m\,j)\, d(j), \qquad m \geq 1.$$

Besides $c *_1 d = c * d$, we also have that $S_a c = a *_2 c$ and we can of course also generalize this to subdivision operator with arbitrary dilation factor, sometimes called "ARITY". This fits nicely into the preceding chapter since

$$\begin{aligned} c *_m d &= \sum_{j\in\mathbb{Z}} c(\cdot - m\,j)\, (\uparrow_m d)\,(m\,j) = \sum_{k\in\mathbb{Z}_m} \sum_{j\in\mathbb{Z}} c(\cdot - m\,j + k) \underbrace{(\uparrow_m d)\,(m\,j+k)}_{=0,\,k\neq 0} \\ &= \sum_{j\in\mathbb{Z}} c(\cdot - j)\, (\uparrow_m d)\,(j) = c * \uparrow_m d \end{aligned}$$

which is indeed a piece of the synthesis part of a filterbank, so that

$$S_a c = a * \uparrow_2 c, \qquad c \in \ell(\mathbb{Z}). \tag{1.2.27}$$

The idea of subdivision is now to iterate $S_a$ on some initial data $c^0 = c \in \ell(\mathbb{Z})$ and to consider $c^n := S_a^n c$ as a function of $2^{-n}\mathbb{Z}$ with denser and denser pixels that eventually converges to a function. The most common definition in this respect looks as follows.

**Definition 1.2.24** (Convergence)**.** The SUBDIVISION SCHEME, i.e., the sequence of operators $S_a^n : \ell(\mathbb{Z}) \to \ell(\mathbb{Z})$, $n \in \mathbb{N}$, with respect to the MASK $a \in \ell_0(\mathbb{Z})$ is said to be a CONVERGENT SUBDIVISION SCHEME if for any $c \in \ell_\infty(\mathbb{Z})$ there exists a uniformly continuous function $f_c : \mathbb{R} \to \mathbb{R}$ such that

$$\lim_{n\to\infty} \sup_{k\in\mathbb{Z}} \left| S_a^n c(k) - f_c\left(2^{-n}k\right)\right| = 0. \tag{1.2.28}$$

Some well–known facts about convergent subdivision schemes are summarized in the next theorem. Since we will not focus on analytic and convergence issues here, we refer, for example, to [Cavaretta et al., 1991, Micchelli, 1995] for the quite elementary proofs.

**Theorem 1.2.25** (Convergence of subdivision)**.** *Let $a \in \ell_0(\mathbb{Z})$ be a given mask.*

1. *The following statements are equivalent:*
   a) *the subdivision scheme based on $a$ converges,*
   b) *there exists a* BASIC LIMIT FUNCTION *$\phi$ such that $S_a^n \delta \to \phi$,*
   c) *$S_a 1 = 1$, where $1 \in \ell(\mathbb{Z})$ stands for the constant sequence, and there exists $b \in \ell_0(\mathbb{Z})$ such that*

   $$a^\sharp(z) = (z+1)\, b^\sharp(z) \tag{1.2.29}$$

   *and $S_b$ is* CONTRACTIVE*, i.e.,*

   $$\lim_{n\to\infty} \left\| S_b^n c\right\|_\infty = 0, \qquad c \in \ell_\infty(\mathbb{Z}). \tag{1.2.30}$$

2. *The basic limit function is* REFINABLE*, i.e.,*

   $$\phi = \sum_{k\in\mathbb{Z}} a(k)\, \phi(2\cdot - k). \tag{1.2.31}$$

3. *Any limit function is a so–called*[12] SEMIDISCRETE CONVOLUTION *of the form*

$$f_c = \sum_{k\in\mathbb{Z}} \phi(\cdot - k)\, c(k) =: \phi * c. \tag{1.2.32}$$

While several of the above results are straightforward consequences of the linearity of the subdivision operator, the convergence result 1c) requires some proof. Note, however, that this is the principle of fixpoint iterations as in Newton's method or the power method in numerical analysis [Gautschi, 1997] and numerical linear algebra [Golub and van Loan, 1996], respectively.

What is more relevant for our purposes is the preservation property $S_a 1 = 1$ and the factorization (1.2.29) – this is algebra and will turn out to be an ideal property in several variables. The connection is made by a simple but important operator.

**Definition 1.2.26** (Difference operator)**.** The DIFFERENCE OPERATOR $\Delta : \ell(\mathbb{Z}) \to \ell(\mathbb{Z})$ is defined as $\Delta = \tau - I$.

**Remark 1.2.27.** Note that the difference operator can also be written as $p(\tau)$ for the polynomial $p(x) = x - 1$, where the variable is *formally* replaced by the shift operator. This algebraization of difference operators will play a fundamental role later.

Since $\delta$ is the unit in the convolution algebra on $\ell(\mathbb{Z})$, we can also write $\Delta c = (\tau\delta - \delta) * c$, $c \in \ell(\mathbb{Z})$ and thus note that the difference operator is indeed a CONVOLUTION OPERATOR. A simple as this is, it means that with respect to symbols or $z$–transforms it is a MULTIPLIER. This is easily verified:

$$(\Delta c)^\flat(z) = \sum_{k\in\mathbb{Z}} (\Delta c)(k)\, z^{-k} = \sum_{k\in\mathbb{Z}} (c(k+1) - c(k))\, z^{-k} = \sum_{k\in\mathbb{Z}} c(k) z^{1-k} - \sum_{k\in\mathbb{Z}} c(k) z^{-k} = (z-1)\, c^\flat(z),$$

hence also

$$(\Delta c)^\sharp(z) = (\Delta c)^\flat(z^{-1}) = (z^{-1} - 1)\, c^\sharp(z).$$

This permanent ambiguity between $z$ and $z^{-1}$ in symbol and $z$–transforms and forward and backward differences[13] is a continuous embarrassment in subdivision and can even be the source for almost religious choices of one of them. In the end it does not matter, one just has to be careful ...

Another important property of the difference is that

$$\Delta c = 0 \qquad \Leftrightarrow \qquad \tau c = c \qquad \Leftrightarrow \qquad c \in \Pi_0, \tag{1.2.33}$$

the KERNEL of the difference operator are exactly the constant sequences.

**Exercise 1.2.3** Show that $\ker \Delta^n = \Pi_{n-1}$, $n \in \mathbb{N}$. ♢

On the other hand, the difference operator is a *minimal annihilator* for arbitrary subdivision operators concerning constants. More precisely.

**Theorem 1.2.28.** *If $m \in \mathbb{N}$ and $a \in \ell_0(\mathbb{Z})$ is such that $0 = S_a 1 := a *_m 1$, then there exists $b \in \ell_0(\mathbb{Z})$ such that $S_a = S_b \Delta$.*

[12] At least by some people.

[13] The BACKWARDS DIFFERENCE is defined as $\tau^{-1} - I$.

**Proof:** Since, for any $a \in \ell_0(\mathbb{Z})$ and $\epsilon \in \mathbb{Z}_m$

$$S_a 1(\epsilon + mk) = \sum_{j\in\mathbb{Z}} a(\epsilon + mk - mj) = \sum_{j\in\mathbb{Z}} a(\epsilon - mj)$$

it follows that

$$S_a 1 = 0 \qquad \Leftrightarrow \qquad \sum_{j\in\mathbb{Z}} a(\epsilon - mj) = 0, \quad \epsilon \in \mathbb{Z}_m.$$

Thus,

$$a^\sharp(1) = \sum_{j\in\mathbb{Z}} a(j) = \sum_{\epsilon\in\mathbb{Z}_m} \sum_{j\in\mathbb{Z}} a(\epsilon - mj) = 0$$

and, for $j \in \mathbb{Z}_m \setminus \{0\}$,

$$\begin{aligned} a^\sharp\left(e^{2\pi ij/m}\right) &= \sum_{k\in\mathbb{Z}} a(k)\left(e^{2\pi ij/m}\right)^k = \sum_{\epsilon\in\mathbb{Z}_m}\sum_{j\in\mathbb{Z}} a(\epsilon - mk)\left(e^{2\pi ij/m}\right)^{\epsilon - mk} \\ &= \sum_{\epsilon\in\mathbb{Z}_m} e^{2\pi i\epsilon j/m} \sum_{j\in\mathbb{Z}} a(\epsilon - mk)\left(e^{-2\pi ij/m}\right)^{mk} = \sum_{\epsilon\in\mathbb{Z}_m} e^{2\pi i\epsilon j/m} \sum_{j\in\mathbb{Z}} a(\epsilon - mk)\underbrace{e^{-2\pi ijk}}_{=1} = 0, \end{aligned}$$

hence

$$a^\sharp\left(e^{2\pi ij/m}\right) = 0, \qquad j \in \mathbb{Z}_m, \tag{1.2.34}$$

and therefore the polynomial

$$\prod_{j\in\mathbb{Z}_m}\left(z - e^{2\pi ij/m}\right) = z^m - 1 \tag{1.2.35}$$

divides $a^\sharp$, hence $a^\sharp(z) = (z^m - 1)\, b^\sharp(z)$ for some $b \in \ell_0(\mathbb{Z})$. On the other hand,

$$(S_b\Delta)^\sharp(z) = (b* \uparrow_m \Delta)^\sharp(z) = b^\sharp(z)\,(z^m - 1) \tag{1.2.36}$$

which completes the proof. □

**Exercise 1.2.4** Prove (1.2.35). If you have no idea how, look up the notion of an *mth root of unity*. ♢

If now $S_a 1 = 1$, then $\Delta S_a 1 = \Delta 1 = 0$ and by Theorem 1.2.28 it follows that there exists $b \in \ell_0(\mathbb{Z})$ such that

$$\Delta S_a = S_b \Delta \tag{1.2.37}$$

or, equivalently,

$$(z-1)\, a^\sharp(z) = (z^m - 1)\, b^\sharp(z) \qquad \Leftrightarrow \qquad a^\sharp(z) = \left(1 + z + \cdots + z^{m-1}\right) b^\sharp(z). \tag{1.2.38}$$

For $m = 2$, this is exactly (1.2.29). These factorizations can be iterated to describe preservation/reproduction of polynomials by a subdivision operator and we will generalize them to several variables, encountering several differences. It will turn out, however, that (1.2.37) extends literally while (1.2.38) requires quite a few new concepts, especially that of a quotient ideal.

### 1.2.4 Prony and moments

PRONY'S PROBLEM is a problem in SPARSE RECONSTRUCTION. The task is to reconstruct a function

$$f(x) = \sum_{j=1}^{n} f_j e^{\omega_j x}, \qquad \omega_j \in \mathbb{R} + i\,\mathbb{T}, \tag{1.2.39}$$

from its values at certain integers. The problem was raised and solved by R. Prony[14] in [Prony, 1795] already in 1795, but in the digital era it became reused, for example in the context of multisource radar, cf. [Schmidt, 1986, Roy and Kailath, 1989]. The standing assumption on the representation (1.2.39) is that it is *efficient* which means that $n$ is the minimal number of terms for the representation[15] of $f$ as a sum of exponentials. This implies that

$$\omega_j \neq \omega_k, \quad j \neq k, \qquad \text{and} \qquad f_j \neq 0, \qquad j = 1, \dots, n. \tag{1.2.40}$$

Moreover, we restrict the imaginary part of the frequencies to be in $\mathbb{T} = \mathbb{R}/2\pi\mathbb{Z} \simeq [-\pi, \pi)$ to avoid ambiguities in the terms $e^{\omega_j}$. In summary, these properties ensure that the representation (1.2.39) is NONREDUNDANT.

**Remark 1.2.29.** Prony's problem actually comes from a real world application in chemistry, namely, the vaporization of alcohol. The function of the form (1.2.39) is a *model* for a liquid that combines several *unknown* vaporization rates of the ingredients[16] that are contained in unknown relative quantities. Or, in a simplified way: *what is the best time to start drinking a glass of whisky*[17].

The interesting point about the problem (1.2.39) is that the determination of the coefficients $f_j$ is a *linear* problem as soon as the FREQUENCIES $\omega_j$ are known, but finding them makes the problem nonlinear.

Prony's trick can even be formulated in terms of digital signal processing: choosing $a \in \ell_0(\mathbb{Z})$, and denoting the restriction of the function $f$ to $\mathbb{Z}$ by $f$ as well, we can compute

$$a * f(j) \quad = \quad \sum_{k\in\mathbb{Z}} a(k) \sum_{\ell=1}^{n} f_\ell e^{\omega_\ell (j-k)} = \sum_{\ell=1}^{n} f_\ell e^{\omega_\ell j} \sum_{k\in\mathbb{Z}} a(k)\, e^{-\omega_\ell k} = \sum_{\ell=1}^{n} f_\ell e^{\omega_\ell j}\, a^\flat\left(e^{\omega_\ell}\right)$$

and write this as

$$f * a = \left(e^{\omega_\ell j} : \begin{array}{c} j \in \mathbb{Z} \\ \ell = 1, \dots, n \end{array}\right) \begin{pmatrix} f_1 & & \\ & \ddots & \\ & & f_n \end{pmatrix} \begin{pmatrix} a^\flat(e^{\omega_1}) \\ \vdots \\ a^\flat(e^{\omega_n}) \end{pmatrix},$$

or, looking at a finite segment only,

$$\begin{pmatrix} (a*f)(0) \\ \vdots \\ (a*f)(n-1) \end{pmatrix} = \begin{pmatrix} 1 & e^{\omega_1} & \dots & (e^{\omega_1})^{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & e^{\omega_n} & \dots & (e^{\omega_n})^{n-1} \end{pmatrix} \begin{pmatrix} f_1 & & \\ & \ddots & \\ & & f_n \end{pmatrix} \begin{pmatrix} a^\flat(e^{\omega_1}) \\ \vdots \\ a^\flat(e^{\omega_n}) \end{pmatrix} \tag{1.2.41}$$

[14] That is the name he uses on the original paper, his full name is *Gaspard Clair François Marie Riche de Prony*

[15] This is the meaning of SPARSITY in this context.

[16] Like alcohol, water and some aromatic content.

[17] There are indeed situations when it becomes necessary to sacrifice oneself in the name of science and to turn even mathematics into an experimental science.

Noting that the first matrix on the right hand side of (1.2.41) is the Vandermonde matrix $V_{n-1}(\{e^{\omega_1},\dots,e^{\omega_n}\})$ and therefore nonsingular as long as the frequencies are disjoint[18], that the middle matrix is diagonal with[19] nonzero diagonal elements, we can draw the following conclusion.

**Proposition 1.2.30.** *A filter with impulse response $a \in \ell_0(\mathbb{Z})$ annihilates $f$ if and only if $a^\flat(e^{\omega_j}) = 0$, $j = 1,\dots,n$.*

This already is Prony's strategy: given the sampled values of $f$, find an ANNIHILATING FILTER $a$ such that $a * f = 0$, which only requires these discrete values, and then compute the zeros of $a^\flat$ to obtain the frequencies. And if we know the "magic" number $n$, we know that there is a unique polynomial of degree $n$ which vanishes at $e^{\omega_j}$, $j = 1,\dots,n$, so if we look for an annihilating filter with only $n+1$ consecutive nonzero "taps", for example by enforcing

$$\operatorname{supp} a := \{k : a(k) \neq 0\} \subset \{0,\dots,n\},$$

then the resulting $a^\flat$ has *exactly* $n$ different zeros and gives the frequencies. To compute this $a$, we note that

$$(a * f)(j) = \sum_{k=0}^{n} a(k) f(j-k), \qquad j = 0,\dots,n-1,$$

so that

$$\begin{pmatrix} (a*f)(0) \\ \vdots \\ (a*f)(n-1) \end{pmatrix} = \begin{pmatrix} f(0) & f(-1) & \dots & f(-n) \\ f(1) & f(0) & \dots & f(1-n) \\ \vdots & \vdots & \ddots & \vdots \\ f(n-1) & f(n-2) & \dots & f(-1) \end{pmatrix} \begin{pmatrix} a(0) \\ \vdots \\ a(n) \end{pmatrix}.$$

Since any nonzero multiple of $a^\flat$ has the same zeros as $a^\flat$, we still have to normalize $a$ which we do by requiring that $a(n) = 1$. Then the normalized annihilating filter is found by solving

$$0 = \begin{pmatrix} (a*f)(0) \\ \vdots \\ (a*f)(n-1) \end{pmatrix} = \begin{pmatrix} f(0) & \dots & f(1-n) \\ \vdots & \ddots & \vdots \\ f(n-1) & \dots & f(0) \end{pmatrix} \begin{pmatrix} a(0) \\ \vdots \\ a(n-1) \end{pmatrix} + \begin{pmatrix} f(-n) \\ \vdots \\ f(-1) \end{pmatrix},$$

hence we solve

$$\underbrace{\begin{pmatrix} f(0) & \dots & f(1-n) \\ \vdots & \ddots & \vdots \\ f(n-1) & \dots & f(0) \end{pmatrix}}_{=:F\in\mathbb{C}^{n\times n}} \underbrace{\begin{pmatrix} a(0) \\ \vdots \\ a(n-1) \end{pmatrix}}_{=:a\in\mathbb{C}^n} = -\underbrace{\begin{pmatrix} f(-n) \\ \vdots \\ f(-1) \end{pmatrix}}_{=:f_n} \tag{1.2.42}$$

and compute the zeros, for example as the eigenvalues of the Frobenius companion matrix

$$\begin{pmatrix} 0 & & & -a(0) \\ 1 & & & -a(1) \\ & \ddots & & \vdots \\ & & 1 & -a(n-1) \end{pmatrix} = \left(e_2,\dots,e_n,F^{-1}f_n\right) = \left(F^{-1}f_1,\dots F^{-1}f_n\right)$$

$$= \begin{pmatrix} f(0) & \dots & f(1-n) \\ \vdots & \ddots & \vdots \\ f(n-1) & \dots & f(0) \end{pmatrix}^{-1} \begin{pmatrix} f(-1) & \dots & f(-n) \\ \vdots & \ddots & \vdots \\ f(n-2) & \dots & f(-1) \end{pmatrix}$$

[18] Which they are according to (1.2.40).
[19] Again due to (1.2.40).

see Example 1.1.22. In particular, the computation of the annihilating filter and its zeros only involves the $2n$ samples $f(-n),\dots,f(n-1)$ which accidentally is precisely the number of unknowns in (1.2.39). This works in principle if and only if the matrix

$$\begin{pmatrix} f(0) & \dots & f(1-n) \\ \vdots & \ddots & \vdots \\ f(n-1) & \dots & f(0) \end{pmatrix}$$

is nonsingular and therefore the solution of (1.2.42) is unique. Since we do not like matrix inverses in numerical computations, we turn everything into the GENERALIZED EIGENVALUE PROBLEM

$$\begin{pmatrix} f(-1) & \dots & f(-n) \\ \vdots & \ddots & \vdots \\ f(n-2) & \dots & f(-1) \end{pmatrix} x = \lambda \begin{pmatrix} f(0) & \dots & f(1-n) \\ \vdots & \ddots & \vdots \\ f(n-1) & \dots & f(0) \end{pmatrix} x \tag{1.2.43}$$

for which there exist numerical methods like the $QZ$ method, described, for example, in [Golub and van Loan, 1996]. What is nice about (1.2.43) is that it computes the frequencies *directly* from the samples of the data without applying any transformations to the problem. There are of course other methods, see for example [Plonka and Tasche, 2014, Potts and Tasche, 2015].

The matrices occurring in (1.2.43) belong to a famous class.

**Definition 1.2.31.** A matrix $A = \big(a_{jk} : j,k\big)$ is called a TOEPLITZ MATRIX if there exists some $a \in \ell(\mathbb{Z})$ such that $a_{jk} = a(j-k)$.

**Remark 1.2.32.** It is easy to see that Toeplitz matrices are the matrix representation of the convolution, seen as a linear operator $T_a$ on $\ell_{\mathbb{Z}}$: $T_a c := a * c$.

**Remark 1.2.33.** The factorization (1.2.41) already hints which concepts we may need in the multivariate case: construction of annihilating filters, Vandermonde matrices, i.e., polynomial interpolation and (common) zeros of polynomials will become the essential tools here, cf. [Sauer, 2017, Sauer, 2018b].

The relationship to moment problems becomes more visible if we write the annihilation concept in a slightly different way by using the CORRELATION

$$c \star d := \sum_{j\in\mathbb{Z}} c(\cdot + k)\, d(k) \tag{1.2.44}$$

Though this operation is no more symmetric, there is no fundamental difference to the convolution and in most application convolutions could be replaced by correlations[20]. Correlations are represented by a different type of matrices.

**Definition 1.2.34.** A matrix $A = \big(a_{jk} : j,k\big)$ is called a HANKEL MATRIX if there exists some $a \in \ell(\mathbb{Z})$ such that $a_{jk} = a(j+k)$.

**Remark 1.2.35.** While Toeplitz matrices are constant on the diagonal and sub- and superdiagonals, Hankel matrices are constant on the so-called *antidiagonals*. For example,

$$H_n = \big(a(j+k) : j,k = 0,\dots,n\big) = \begin{pmatrix} f(0) & f(1) & \dots & f(n) \\ f(1) & f(2) & \dots & f(n+1) \\ \vdots & \vdots & \ddots & \vdots \\ f(n) & f(n+1) & \dots & f(2n) \end{pmatrix}$$

is such a typical Hankel matrix.

[20] This is comparable to the choice between $z$-transforms and symbol.

Hankel matrices are classical objects in mathematics, in particular due to their immediate relationship to the MOMENT PROBLEM.

**Definition 1.2.36.** Given a measure $\mu$ on $\mathbb{R}$, the $j$th MOMENT of $\mu$ is defined as

$$\mu(j) := \int_{\mathbb{R}} x^j \, d\mu(x), \qquad j \in \mathbb{N}_0. \tag{1.2.45}$$

The sequence $\big(\mu(j) : j \in \mathbb{N}_0\big)$ is called the MOMENT SEQUENCE and moment problems consider the question which sequences are moments for certain types of measures. Moreover, the FOURIER TRANSFORM of the measure $\mu$ is defined as

$$\hat{\mu}(\xi) := \int_{\mathbb{R}} e^{-ix\xi} \, d\mu(x), \qquad \xi \in \mathbb{R}. \tag{1.2.46}$$

Since the moment matrix

$$M = \left( \int_{\mathbb{R}} x^j \, x^k \, d\mu(x) : j,k \right) = \big(\mu(j+k) : j,k\big)$$

is a Hankel matrix that describes the action of the linear functional $f \mapsto \int f \, d\mu$ on polynomials, Hankel matrices and Hankel operators naturally connect to moment problems. In the special case of a signed DISCRETE MEASURE of the form

$$\mu = \sum_{j=1}^{n} f_j \, \delta_{i\omega_j}, \qquad f_j, \omega_j \in \mathbb{C},$$

whose Fourier transform takes the familiar form

$$\hat{\mu}(\xi) = \sum_{j=1}^{n} f_j \int_{\mathbb{R}} e^{-i\xi x} \, d\delta_{i\omega_j}(x) = \sum_{j=1}^{n} f_j \, e^{\omega_j \xi},$$

the task of reconstructing the locations and weights of the measure is then exactly Prony's problem where we reconstruct from the sequence

$$\hat{\mu}(k) = \sum_{j=1}^{n} f_j \left( e^{-i\omega_j} \right)^k = \sum_{j=1}^{n} f_j \int_{\mathbb{R}} x^k \, d\delta_{e^{-i\omega_j}}(x) = \tilde{\mu}(j)$$

which is the moment sequence of the discrete measure

$$\tilde{\mu} = \sum_{j=1}^{n} f_j \delta_{e^{-i\omega_j}}$$

with the same weights and re-localized centers of mass.

In summary, Prony's problem, finite discrete signed measures, Hankel or Toeplitz operators of finite rank are all just different points of view for the same thing. This also holds true in several variables and will enable us to combine all the theory that we learn in the sequel.

# Constructive ideal theory | 2

*Men invent new ideals because they dare not attempt old ideals. They look forward with enthusiasm, because they are afraid to look back*

(G. K. Chesterton, *What's wrong with the world*)

In this chapter we make a systematic approach to the techniques needed for dealing with ideals of multivariate polynomials. We will denote by $\mathbb{K}$ a field of characteristic zero, but mainly have the three cases $\mathbb{Q}, \mathbb{R}, \mathbb{C}$ in mind. We will write $\mathbb{K}_\times = \mathbb{K} \setminus \{0\}$ for the units of $\mathbb{K}$.

## 2.1 Polynomial and Laurent ideals

We begin with carefully defining the necessary concepts for dealing with *multivariate* polynomials. This needs some notation and terminology, in particular if we still want to write things in a compact way. In all that we are doing, we will use $s \in \mathbb{N}$ for the number of variables[1], so that polynomials can always be seen as functions $\mathbb{K}^s \to \mathbb{K}$.

### 2.1.1 (Laurent) polynomials in several variables

We begin with some standard notation for multivariate objects.

**Definition 2.1.1** (Multiindices, monomials and terms)**.**

1. A MULTIINDEX $\alpha = (\alpha_1, \ldots, \alpha_s) \in \mathbb{Z}^s$ is a tuple of integers. Its LENGTH is defined as $|\alpha| = |\alpha_1| + \cdots + |\alpha_s|$. For $\alpha \in \mathbb{N}_0^s$, we also define its FACTORIAL as $\alpha! := \alpha_1! \cdots \alpha_s!$. Moreover, we use the partial ordering $\alpha \le \beta$ if $\alpha_j \le \beta_j$, $j = 1, \ldots, s$.

2. A MONOMIAL in $x \in \mathbb{K}^s$ is an expression

$$x^\alpha := x_1^{\alpha_1} \cdots x_s^{\alpha_s}, \qquad \alpha \in \mathbb{N}_0^s \text{ or } \alpha \in \mathbb{Z}^s. \tag{2.1.1}$$

   If we want to emphasize that we permit an arbitrary $\alpha \in \mathbb{Z}^s$, we will speak of a LAURENT MONOMIAL. A TERM is an expression of the form $c\, x^\alpha$, $c \in \mathbb{K}_\times$.

3. A POLYNOMIAL is a *finite* linear combination of monomials over $\mathbb{K}$,

$$f(x) = \sum_{\alpha \in \mathbb{N}_0^s} f_\alpha\, x^\alpha, \qquad \#\{\alpha : f_\alpha \neq 0\} < \infty, \tag{2.1.2}$$

[1] The more common use in algebra is to use $n$ for the number of variables and $d$ for the **d**egree, in analysis people often prefer to use $d$ for the space **d**imension and $n$ for the degree. This can be confusing sometimes. The choice here is just a personal **s**election.

a LAURENT POLYNOMIAL a finite linear combination of Laurent monomials,

$$f(x) = \sum_{\alpha\in\mathbb{Z}^s} f_\alpha\, x^\alpha, \qquad \#\{\alpha : f_\alpha \neq 0\} < \infty. \tag{2.1.3}$$

The ring[2] of all polynomials is written as $\Pi = \mathbb{K}[x]$, the ring of Laurent polynomials as $\Lambda$.

4. The SUPPORT of a (Laurent) polynomial is the set of the indices of nonzero coefficients:

$$\operatorname{supp} f := \{\alpha \in \mathbb{Z}^s : f_\alpha \neq 0\}. \tag{2.1.4}$$

It is worthwhile to record some simple and elementary consequences of these definitions.

**Remark 2.1.2** (On Definition 2.1.1)**.**

1. The zero polynomial is not considered to be a term.
2. Laurent monomials are only defined on $\mathbb{K}^s_\times$.
3. Any polynomial is a Laurent polynomial and any Laurent polynomial can be written as

$$f(x) = x^\alpha\, p(x), \qquad \alpha \in \mathbb{Z}^s,\ p \in \Pi.$$

This representation is not unique, but can be made unique by requesting that $p(0) \neq 0$, hence the support of $p$ cannot be shifted any further and $\alpha$ is then a maximal choice since

$$f(x) = x^{\alpha-\beta}\, \underbrace{x^\beta p(x)}_{\in\Pi}, \qquad \beta \in \mathbb{N}_0^s,$$

would be a valid representation as well.

4. Polynomials are expressions with finite support in $\mathbb{N}_0^s$, Laurent polynomials expressions with finite support in $\mathbb{Z}^s$.

**Definition 2.1.3** (Differential operators)**.** Given a polynomial $f \in \Pi$, the associated PARTIAL DIFFERENTIAL OPERATOR $f(D)$ is obtained by formally replacing $x$ by the differential $\frac{\partial}{\partial x}$, that is,

$$f(D) = \sum_{\alpha\in\mathbb{N}_0^s} f_\alpha \frac{\partial^\alpha}{\partial x^\alpha}, \qquad f = \sum_{\alpha\in\mathbb{N}_0^s} f_\alpha\, (\cdot)^\alpha. \tag{2.1.5}$$

**Definition 2.1.4.** An INNER PRODUCT $(\cdot,\cdot)$ on $\Pi$ is, in the case of $\mathbb{K} \subseteq \mathbb{R}$ a symmetric, definite bilinear form, for $\mathbb{C}$ only a definite sesquilinear form, that is

$$(f,g) = \overline{(g,f)}, \qquad (f,f) \in \mathbb{R}_+.$$

**Example 2.1.5.** The simplest way to define an inner product would be to just take the inner product of the coefficients,

$$(f,g) = \sum_{\alpha\in\mathbb{N}_0^s} f_\alpha\, g_\alpha, \tag{2.1.6}$$

but a nicer one is

$$(f,g) = \big(f(D)g\big)(0) = \sum_{\alpha\in\mathbb{N}_0^s} \alpha!\, f_\alpha\, g_\alpha. \tag{2.1.7}$$

[2] This means that we will not only add and multiply by field elements as in a vector space, but also multiply polynomials with each other, following Gen 1:28.

**Exercise 2.1.1** Prove that the two bilinear forms in (2.1.6) and (2.1.7) are inner products in $\mathbb{R}[x]$ and verify (2.1.7). Deriver their extensions to $\mathbb{C}[x]$. ♢

**Remark 2.1.6.** The inner product from (2.1.7) is, among others, known as the FISHER INNER PRODUCT or BOMBIERI INNER PRODUCT and has been used in various instances, for example also by Gröbner in [Gröbner, 1937].

The inner product (2.1.7) has another property that allows us get an explicit Riesz representation of point evaluations when extending the inner product properly.

**Definition 2.1.7** (Power series)**.** The algebra $\mathbb{K}(x)$ of FORMAL POWER SERIES over $\mathbb{K}$ consists of all expressions of the form

$$f(x) = \sum_{\alpha\in\mathbb{N}_0^s} f_\alpha\, x^\alpha, \qquad f_\alpha \in \mathbb{K}. \tag{2.1.8}$$

A sequence $f_n$, $n \in \mathbb{N}$, is said to be CONVERGENT to $f \in \mathbb{K}(x)$ if for any finite $\Omega \subset \mathbb{N}_0^s$ there exists an $n_0 \in \mathbb{N}$ such that

$$f_{n,\alpha} = f_\alpha, \qquad \alpha \in \Omega, \qquad n \geq n_0. \tag{2.1.9}$$

For $y \in \mathbb{K}^s$ we define the element $e_y \in \mathbb{K}(x)$ as

$$e_y(x) := \sum_{\alpha\in\mathbb{N}_0^s} \frac{y^\alpha}{\alpha!}\, x^\alpha. \tag{2.1.10}$$

Recall the classical fact[3] that the product of two power series

$$\left(\sum_{\alpha\in\mathbb{N}_0^s} f_\alpha\, x^\alpha\right)\left(\sum_{\alpha\in\mathbb{N}_0^s} g_\alpha\, x^\alpha\right) = \sum_{\alpha,\beta\in\mathbb{N}_0^s} f_\alpha g_\beta\, x^{\alpha+\beta} = \sum_{\alpha\in\mathbb{N}_0^s}\left(\sum_{0\leq\beta\leq\alpha} f_{\alpha-\beta} g_\beta\right) x^\alpha$$

relies on the CAUCHY PRODUCT of the coefficients which is indeed a convolution and computes any coefficient of the product from only *finitely many* coefficients, hence the product is well–defined in $\mathbb{K}(x)$.

Since the product in (2.1.7) is well–defined when $f$ or $g$ has finite support, we can extend it to $\mathbb{K}[x] \times \mathbb{K}(x)$ and have the following simple but extremely useful result that has been used in many applications, cf. [Boor and Ron, 1992, Mourrain, 2016].

**Theorem 2.1.8.** *For any $f \in \Pi$ and $x \in \mathbb{K}$, we have that*

$$f(x) = (f, e_x). \tag{2.1.11}$$

**Proof:** We only have to note that, due to (2.1.7),

$$(f, e_x) = \sum_{\alpha\in\mathbb{N}_0} \alpha! f_\alpha \frac{x^\alpha}{\alpha!} = \sum_{\alpha\in\mathbb{N}_0} f_\alpha\, x^\alpha = f(x)$$

to verify (2.1.11). □

[3] In one variable it is part in any calculus lecture.

### 2.1.2 Ideals and varieties

Now we get to the fundamental concept of this lecture.

**Definition 2.1.9** (Ideal)**.** An IDEAL $\mathscr{I}$ in a ring $R$ is a subset that is closed under addition and under multiplication with arbitrary elements from $R$, i.e., $\mathscr{I}+\mathscr{I}=\mathscr{I}$, $\mathscr{I}\cdot R=\mathscr{I}$. A TRIVIAL IDEAL is either $\mathscr{I}=\{0\}$ and $\mathscr{I}=R$. An ideal that does not equal $R$ is called PROPER.

Trivial ideals can be identified easily: they either consist only of 0 or they contain units.

**Lemma 2.1.10.** *An ideal $\mathscr{I}\subset R$ satisfies $\mathscr{I}=R$ if and only if $\mathscr{I}\cap R^\times\neq\emptyset$.*

**Proof:** The direction "⇒" is trivial and if $r\in\mathscr{I}$ for some $r\in R^\times$ then

$$a=ar^{-1}r\in r\cdot R=\mathscr{I},\qquad a\in R,$$

hence $\mathscr{I}=R$. □

**Corollary 2.1.11.** *An ideal $\mathscr{I}\subset\mathbb{K}[x]$ is trivial in the sense $\mathscr{I}=\mathbb{K}[x]$ if and only if $1\in\mathscr{I}$.*

**Remark 2.1.12.** If the ring $R$ contains a field $\mathbb{K}$, as our (Laurent) polynomials do, then any ideal is also a $\mathbb{K}$ vector space; nonlinearity enters due to the *multiplication* of polynomials.

**Remark 2.1.13.** If $\mathscr{I}$ and $\mathscr{J}$ are ideals, then $\mathscr{I}\cap\mathscr{J}$ is an ideal as well as the defining closedness conditions are preserved individually.

In the sequel we will only consider ideals in $\Pi$ and $\Lambda$ and forget about more general rings. The first observation is that there are two simple ways to generate ideals.

**Example 2.1.14** (Ideal constructions)**.**

1. If $F$ is a finite subset of $\Pi$ or $\Lambda$, then its COMPLETION with respect to ideal operations,

$$\langle F\rangle_\Pi:=\left\{\sum_{f\in F}g_f f : g_f\in\Pi\right\},\qquad \langle F\rangle_\Lambda:=\left\{\sum_{f\in F}g_f f : g_f\in\Lambda\right\},\tag{2.1.12}$$

   are ideals in $\Pi$ and $\Lambda$, respectively, the ideals GENERATED BY $F$. Note, however, that for $F\subset\Pi$ the ideals $\langle F\rangle_\Pi$ and $\langle F\rangle_\Lambda$ are clearly different. Normally, we will drop the subscript in (2.1.12) if it is clear in which of the two rings the ideal is formed.

2. If $\mathscr{X}\subset\mathbb{K}^s$, then the set

$$I(\mathscr{X}):=\mathscr{I}_{\mathscr{X}}:=\left\{f : f(\mathscr{X})=\{0\}\right\}\tag{2.1.13}$$

   of all elements VANISHING at $\mathscr{X}$ is an ideal as well, called the ZERO IDEAL[4] of $\mathscr{X}$. Note, that in the case of ideals in $\Lambda$ we have to require that $\mathscr{X}\subset\mathbb{K}^s_\times$.

The first simple observation is that the solution of a SYSTEM OF EQUATIONS $F(x)=0$, $F\subset\Pi$, is a matter of the ideal $\langle F\rangle$ and not of the specific equations $F$. Indeed,

$$f(x)=0,\quad f\in F\qquad\Rightarrow\qquad\sum_{f\in F}g_f(x)\,f(x)=0,\quad g_f\in\Pi,\qquad\Rightarrow\qquad f(x)=0,\quad f\in\langle F\rangle,$$

and since the converse is trivial due to $F\subseteq\langle F\rangle$, we can conclude that

$$\mathscr{F}(x)=0\qquad\Leftrightarrow\qquad\langle F\rangle(x)=0.\tag{2.1.14}$$

This is the reason why the next definition restricts to ideals without any loss of generality.

[4] The German terminology NULLSTELLENINDEAL is somewhat nicer.

**Definition 2.1.15.** For an ideal $\mathscr{I} \subset \Pi$ the associated VARIETY is defined as

$$V(\mathscr{I}) := \left\{x \in \mathbb{K}^s : f(x) = 0,\, f \in \mathscr{I}\right\}, \tag{2.1.15}$$

and for $\mathscr{I} \subset \Lambda$,

$$V_\Lambda(\mathscr{I}) := \left\{x \in \mathbb{K}_\times^s : f(x) = 0,\, f \in \mathscr{I}\right\}. \tag{2.1.16}$$

The connection between *polynomial* ideals and varieties is expressed in the following result whose proof can be found in [Cox et al., 1996, p. 168–171]. So we first focus on polynomial ideals and point out similarities and differences later in Section 2.1.6, where we will also show how to reduce Laurent ideals to polynomial ideals

**Theorem 2.1.16** (NULLSTELLENSATZ)**.** *Let $\mathbb{K}$ be an algebraically closed field*[5].

1. *If $\mathscr{I} \subset \mathbb{K}[x]$ is an ideal with $V(\mathscr{I}) = \emptyset$, then $\mathscr{I} = \mathbb{K}[x]$.*
2. *If $f \in I(V(\mathscr{I}))$ then there exists some $m \in \mathbb{N}$ such that $f^m \in \mathscr{I}$.*

**Remark 2.1.17.** Theorem 2.1.16 gives two versions of the famous Nullstellensatz. The statement 1) is usually called the WEAK NULLSTELLENSATZ and proved by induction on the number of variables in which way it can be seen the generalization of the fact that the only univariate polynomials over $\mathbb{C}$ that have no zero must be constant. Statement 2) is known as HILBERT'S NULLSTELLENSATZ and already addresses some of the problems caused by multiplicity of zeros.

### 2.1.3 Simple ideal operations

There are elementary operations that we can apply to an ideal. They are defined as follows.

**Definition 2.1.18** (Ideal operations)**.** Let $\mathscr{I}$, $\mathscr{J}$ be two ideals.

1. The sum and the product are defined as

$$\mathscr{I} + \mathscr{J} := \left\{f + g : f \in \mathscr{I}, g \in \mathscr{J}\right\}, \qquad \mathscr{J} \cdot \mathscr{J} := \left\langle fg : f \in \mathscr{I}, g \in \mathscr{J}\right\rangle. \tag{2.1.17}$$

2. The QUOTIENT IDEAL $\mathscr{I} : \mathscr{J}$ is defined as

$$\mathscr{I} : \mathscr{J} := \left\{f \in \Pi : f \cdot \mathscr{J} \subseteq \mathscr{I}\right\} \tag{2.1.18}$$

**Proposition 2.1.19.** *$\mathscr{I} + \mathscr{J}$, $\mathscr{I}\,\mathscr{J}$ and $\mathscr{I} : \mathscr{J}$ are ideals and satisfy*

$$\mathscr{I} \cdot \mathscr{J} \subseteq \mathscr{I},\, \mathscr{J} \subseteq \mathscr{I} + \mathscr{J}, \qquad \mathscr{I} : \mathscr{J} \supseteq \mathscr{I}. \tag{2.1.19}$$

**Proof:** The $\mathscr{I} + \mathscr{J}$ is an ideal follows for $f, f' \in \mathscr{I}$, $g, g' \in \mathscr{J}$ from

$$(f + g) + (f' + g') = (f + f') + (g + g'), \qquad p(f + g) = pf + pg \in \mathscr{I} + \mathscr{J}$$

and is trivial for $\mathscr{I}\,\mathscr{J}$. Also the inclusions are immediate. The quotient ideal is an ideal since for $f, f' \in \mathscr{I} : \mathscr{J}$

$$(f + f')\,\mathscr{J} = \{(f + f')g : g \in \mathscr{J}\} \subseteq \{fg + f'g' : g, g' \in \mathscr{J}\} = f \cdot \mathscr{J} + f' \cdot \mathscr{J} \subseteq \mathscr{I} + \mathscr{I} = \mathscr{I}$$

and, for $p \in \Pi$,

$$pf\,\mathscr{J} = f(p\,\mathscr{J}) \subseteq f\,\mathscr{J} \subseteq \mathscr{I}.$$

Hence, it is an ideal, the inclusion follows from $\mathscr{I} \cdot \mathscr{J} \subseteq \mathscr{J}$. □

[5] In particular, $\mathbb{K} = \mathbb{C}$.

### 2.1.4 Ideal types: from radical to primary

In view of Theorem 2.1.16, we define some more terminology on polynomial ideals.

**Definition 2.1.20.** An ideal $\mathscr{I}$ is called

1. RADICAL IDEAL if $\mathscr{I} = \sqrt{\mathscr{I}}$ where the RADICAL of $\mathscr{I}$ is defined as
$$\sqrt{\mathscr{I}} = \{f \in \Pi : f^m \in \mathscr{I} \text{ for some } m\}. \tag{2.1.20}$$
2. MAXIMAL IDEAL if $\mathscr{I} \neq \Pi$ and $\mathscr{I} \subseteq \mathscr{J}$ for some ideal $\mathscr{J}$ implies $\mathscr{J} = \mathscr{I}$ or $\mathscr{J} = \Pi$.
3. PRIME IDEAL if $fg \in \mathscr{I}$ implies either $f \in \mathscr{I}$ or $g \in \mathscr{I}$.
4. PRIMARY IDEAL if $fg \in \mathscr{I}$ implies either $f \in \mathscr{I}$ or $g^m \in \mathscr{I}$ for some $m \geq 1$.

**Remark 2.1.21.** Some remarks concerning the concepts from Definition 2.1.20:

1. By definition we have that $\mathscr{I} \subseteq \sqrt{\mathscr{I}}$.
2. Maximal ideals also play a fundamental role in functional analysis, namely, in the context of BANACH ALGEBRAS, cf. [Yosida, 1965].
3. The relationship between primary and prime ideals is a radical one. Indeed, see [Cox et al., 1996, p. 207], for any primary ideal $\mathscr{I}$ its radical $\sqrt{\mathscr{I}}$ is prime and is indeed the smallest prime ideal containing $\mathscr{I}$.

In terms of these definitions we can rephrase Statement 2 of Theorem 2.1.16 as follows.

**Corollary 2.1.22** (STRONG NULLSTELLENSATZ)**.** *If $\mathbb{K}$ is algebraically closed then $I(V(\mathscr{I})) = \sqrt{\mathscr{I}}$ for any ideal in $\mathbb{K}[x]$.*

**Proposition 2.1.23.** *For any variety $V \subseteq \mathbb{K}^s$, the ideal $I(V)$ is radical.*

**Proof:** For $f \in \sqrt{I(V)}$ we have $0 = f^m(x)$, $x \in V$, i.e., $0 = f(x)$, $x \in V$, hence $f \in I(V)$. This shows that $\sqrt{I(V)} \subseteq I(V)$ and since the converse inclusion is true by definition, the two ideals coincide. □

Also maximal and prime ideals share a close relationship.

**Proposition 2.1.24.** *Let $\mathscr{I} \subset \mathbb{K}[x]$ be a maximal ideal.*

1. *$\mathscr{I}$ is prime.*
2. *If $\mathbb{K}$ is algebraically closed, then there exists $z = (z_1, \ldots, z_s) \in \mathbb{K}^s$ such that*
$$\mathscr{I} = \langle (\cdot) - z \rangle := \langle (\cdot)_j - z_j : j = 1, \ldots, s \rangle. \tag{2.1.21}$$

**Proof:** For 1) suppose that $\mathscr{I}$ is not prime, hence here exist $f, g \notin \mathscr{I}$ such that $fg \in \mathscr{I}$. Then $\mathscr{I} \subset \langle f \rangle + \mathscr{I}$ since the ideal on the right hand side contains $f$ which does not belong to $\mathscr{I}$. On the other hand, $\langle f \rangle + \mathscr{I} = \mathbb{K}[x]$ would imply that $1 \in \langle f \rangle + \mathscr{I}$, that is, there exist $h \in \mathscr{I}$ and $p \in \mathbb{K}[x]$ such that $1 = pf + h$, hence
$$g = g \cdot 1 = g(pf + h) = p \underbrace{gf}_{\in \mathscr{I}} + g \underbrace{h}_{\in \mathscr{I}} \in \mathscr{I},$$

which is a contradiction. Consequently, $\mathscr{I}$ is prime as claimed.

For statement 2) we note that, by the Nullstellensatz $\mathscr{I} \neq \Pi$ implies that $V(\mathscr{I}) \neq \emptyset$, hence there exists some $z \in V(\mathscr{I})$. This implies together with Theorem 2.1.22 that

$$I(z) \supseteq I(V(\mathscr{I})) = \sqrt{\mathscr{I}} = \mathscr{I}$$

since, by 1), $\mathscr{I}$ is prime, hence equals the smallest prime ideal containing $\mathscr{I}$ which is $\sqrt{\mathscr{I}}$, see Remark 2.1.21. Therefore,

$$\mathscr{I} \subseteq I(z) = \langle (\cdot)_j - z_j : j = 1, \ldots, s \rangle \subset \mathbb{K}[x]$$

and maximality of $\mathscr{I}$ finally implies that $\mathscr{I} = I(z)$. □

**Exercise 2.1.2** Show that

$$I(z) = \langle (\cdot)_j - z_j : j = 1, \ldots, s \rangle, \qquad z \in \mathbb{K},$$

holds for any algebraically closed field. ♢

Primary ideals form the building blocks of polynomial ideals. In fact, *any* ideal $\mathscr{I}$ in $\mathbb{K}[x]$ can be written as a *finite* intersection of primary ideals. There is an even more advanced version of it, the LASKER-NOETHER THEOREM that describes the situation in detail. It can be found, with its proof[6] in [Cox et al., 1996, Theorem 7 and 9, p. 208], which we repeat in a summarized form.

**Theorem 2.1.25** (Lasker–Noether)**.** *Every polynomial ideal $\mathscr{I} \subseteq \mathbb{K}[x]$ has a finite* minimal *primary decomposition*

$$\mathscr{I} = \bigcap_{j=1}^{n} \mathscr{I}_j, \qquad \mathscr{I}_j \text{ primary}, \tag{2.1.22}$$

*where*[7] *$\sqrt{\mathscr{I}_j}$ are distinct and $\mathscr{I}_j \not\supset \bigcap_{k \neq j} \mathscr{I}_k$. Moreover, the proper prime ideals among $\sqrt{\mathscr{I} : \langle f \rangle}$, $f \in \mathbb{K}[x]$, are exactly $\sqrt{\mathscr{I}_j}$, $j = 1, \ldots, n$.*

Dimension theory of ideals in general is a nontrivial issue, see [Cox et al., 1996] and [Gröbner, 1970], but one special case is easy to describe already at this point. Since it is the most relevant one in our applications, we give the definition here.

**Definition 2.1.26.** $\mathscr{I} \subseteq \mathbb{K}[x]$ with $\mathbb{K}$ algebraically closed is called a ZERO DIMENSIONAL IDEAL if the associated variety is *finite*: $\#V(\mathscr{I}) < \infty$.

An immediate consequence of the primary decomposition from the Lasker-Noether Theorem is as follows.

**Theorem 2.1.27.** *Any zero dimensional ideal has a primary decomposition of the form*

$$\mathscr{I} = \bigcap_{z \in V(\mathscr{I})} \langle \cdot - z \rangle^{k_z}, \qquad k_z \in \mathbb{N}, z \in V(\mathscr{I}). \tag{2.1.23}$$

[6] The proof is quite elementary but the necessary details would distract us too far from what we want to do here.

[7] This is the definition of *minimal*.

The intuition behind the proof of Theorem 2.1.27 is that $\sqrt{\mathscr{I}}$ is a zero dimensional prime ideal that can be decomposed into its maximal parts

$$\sqrt{\mathscr{I}} = \bigcap_{z \in V(\mathscr{I})} \langle (\cdot) - z \rangle$$

and then the "local multiplicities" are handled by passing to the primary parts. We will not give a detailed proof here as we will consider more precise statements with exact multiplicity of zeros later on.

### 2.1.5 Bases

The next definition is fundamental for all the theory we deal with later.

**Definition 2.1.28** (Ideal basis)**.** A set $F \subset \Pi$ or $F \subset \Lambda$ is called a basis for an ideal $\mathscr{I} \subset \Pi$ or $\mathscr{I} \subset \Lambda$ if $\mathscr{I} = \langle F \rangle_\Pi$ or $\mathscr{I} = \langle F \rangle_\Lambda$, respectively. If the context is clear, we simply write $\mathscr{I} = \langle F \rangle$.

The following result ensures that ideals can be handled on a computer and justifies the existence of computer algebra systems like `Maple` or `Mathematica`. It is Hilbert's famous Basissatz.

**Theorem 2.1.29** (BASISSATZ)**.** *Any ideal in* $\Pi$ *has a* ***finite*** *basis.*

The proof of this fundamental result can be found in all books on algebraic geometry, for example in [Gröbner, 1968], but also in [Cox et al., 1996, Eisenbud, 1994]. For our purposes so far it suffices to believe it, following the famous quote

> *The proof of the Hilbert Basis Theorem is not mathematics; it is theology.*
>
> (P. Gordan)[8]

Once we represent ideals by means of their **finite** bases, we have to figure out whether and how we can perform ideal operations by means of the bases; later we will also determine how this can be done *efficiently*. In what follows, we consider the case of two ideals

$$\mathscr{I} = \langle F \rangle, \quad \mathscr{J} = \langle G \rangle, \qquad F, G \subset \mathbb{K}[x], \qquad \#F, \#G < \infty. \tag{2.1.24}$$

Addition is indeed simple.

**Lemma 2.1.30.** $\mathscr{I} + \mathscr{J} = \langle F \cup G \rangle$.

**Proof:** For $p \in \mathscr{I}$ and $q \in \mathscr{J}$ we write

$$p = \sum_{f \in F} p_f f \qquad \text{and} \qquad q = \sum_{g \in F} q_g g,$$

and get

$$p + q = \sum_{f \in F} p_f f + \sum_{g \in F} q_g g = \sum_{h \in F \cup G} p_h h.$$

□

[8] In [Eisenbud, 1994] this statement is attributed to the "reigning king of invariants" who happened to work in Gießen and Erlangen, but not in Passau. In the online Math Tutor [MacTutor, 2003], it can even be found twice, as due to Paul Gordan and Camille Jordan, but Gordan seems to be more likely.

**Lemma 2.1.31.** $\mathscr{I} \cdot \mathscr{J} = \langle f g : f \in F, g \in G \rangle$.

**Proof:** By definition, an arbitrary element of $\mathscr{I} \cdot \mathscr{J}$ is of the form

$$a = \sum_{j=1}^{n} b_j \, p_j \, q_j, \qquad b_j \in \Pi, \, p_j \in \mathscr{I}, \, q_j \in \mathscr{J}, \qquad j = 1, \ldots, n,$$

and the respective bases yield

$$a = \sum_{j=1}^{n} b_j \left( \sum_{f \in F} p_{jf} \, f \right) \left( \sum_{g \in G} q_{jg} \, g \right) = \sum_{f \in F} \sum_{g \in G} \left( \sum_{j=1}^{n} b_j \, p_{jf} \, q_{jg} \right) f g,$$

which proves the claim. □

The IDEAL INTERSECTION is more complex and based on the following result.

**Lemma 2.1.32.** *In* $\mathbb{K}[x,t]$ *one has*

$$\mathscr{I} \cap \mathscr{J} = \left( t \, \mathscr{I} + (1-t) \, \mathscr{J} \right) \cap \mathbb{K}[x]. \tag{2.1.25}$$

*Hence, any basis of* $\langle tF + (1-t)G \rangle \cap \mathbb{K}[x]$ *is a basis of* $\mathscr{I} \cap \mathscr{J}$.

**Proof:** We denote the ideal on the right hand side of (2.1.25) by $\mathscr{H}$. For $f \in \mathscr{I} \cap \mathscr{J}$, we have the trivial identity

$$f = (t + (1-t)) f = t f + (1-t) f \in t.\mathscr{I} + (1-t) \, \mathscr{J} = \mathscr{H}$$

hence $\mathscr{I} \cap \mathscr{J} \subseteq \mathscr{H}$. If, on the other hand, we can write $p \in \mathbb{K}[x,t]$ as

$$p(x,t) = t f(x) + (1-t) g(x)$$

then

$$p(x,0) = 0 \, f(x) + (1-0) \, g(x) = g(x) \in \mathscr{J},$$

and

$$p(x,1) = 1 \, f(x) + (1-1) \, g(x) = f(x) \in \mathscr{I},$$

and if $p \in \mathscr{H}$, i.e., $p(x,t) = p(x)$, is independent of $t$, then $p(x) \in \mathscr{I} \cap \mathscr{J}$, hence $\mathscr{I} \cap \mathscr{J} \supseteq \mathscr{H}$. □

To really *compute* the ideal intersection, we need to be able to determine a basis of $\langle tF + (1-t)G \rangle$ with $t$ being eliminated. Such an elimination ideal can be computed by means of Gröbner bases.

The last step is how to determine a basis for the quotient ideal which will be based on the ideal intersection and relies on the following observations that enable us to compute the basis step by step.

**Theorem 2.1.33** (Properties of quotient ideals)**.**

1. *For ideals* $\mathscr{I}$, $\mathscr{J}$, $\mathscr{J}'$ *the following holds:*

$$\mathscr{I} : (\mathscr{J} + \mathscr{J}') = (\mathscr{I} : \mathscr{J}) \cap (\mathscr{I} : \mathscr{J}'). \tag{2.1.26}$$

2. *If $F$ is a basis of $\mathscr{I} \cap \langle g \rangle$, then*

$$F/g := \left\{ \frac{f}{g} : f \in F \right\} \tag{2.1.27}$$

*is a basis of $\mathscr{I} : \langle g \rangle$.*

**Proof:** For 1) we first observe that $f \in \mathscr{I} : (\mathscr{J} + \mathscr{J}')$ means

$$f(g+g') \in \mathscr{I}, \qquad g \in \mathscr{J}, g' \in \mathscr{J}',$$

which holds in particular for $g = 0$ or $g' = 0$, hence

$$fg \in \mathscr{I}, fg' \in \mathscr{I}, \quad g \in \mathscr{J}, g' \in \mathscr{J}' \qquad \Rightarrow \qquad f \in (\mathscr{I} : \mathscr{J}) \cap (\mathscr{I} : \mathscr{J}'),$$

and the inclusion "'$\subseteq$"' in (2.1.26) is verified. If, conversely, $f \in (\mathscr{I} : \mathscr{J}) \cap (\mathscr{I} : \mathscr{J}')$, then

$$f\mathscr{J} \subseteq \mathscr{I}, f\mathscr{J}' \subseteq \mathscr{I} \qquad \Rightarrow \qquad f(\mathscr{J} + \mathscr{J}') \subseteq \mathscr{I},$$

gives "'$\supseteq$"' and completes the proof of (2.1.26).

For $p \in \langle F/g \rangle$ and $q = hg \in \langle g \rangle$ we have that

$$pq = hg \sum_{f \in F} p_f \frac{f}{g} = \sum_{f \in F} (hp_f) f \in \langle F \rangle = \mathscr{I} \cap \langle g \rangle \subseteq \mathscr{I},$$

hence $p \in \mathscr{I} : \langle g \rangle$, that is, $\langle F/g \rangle \subseteq \mathscr{I} : \langle g \rangle$. For the converse we assume that $p \in \mathscr{I} : \langle g \rangle$, yielding, in particular, that $pg \in \mathscr{I}$. On the other hand $pg \in \langle g \rangle$, implying $pg \in I \cap \langle g \rangle = \langle F \rangle$ or

$$pg = \sum_{f \in F} q_f f \qquad \Rightarrow \qquad p = \sum_{f \in F} q_f \frac{f}{g} \in \langle F/g \rangle. \tag{2.1.28}$$

Since $F \subset \mathscr{I} \cap \langle g \rangle$, it follows that

$$f \in F \qquad \Rightarrow \qquad f \in \langle g \rangle \qquad \Rightarrow \qquad f = g_f g$$

an therefore the quotients $f/g$ in (2.1.28) are really polynomials[9]. In total we have shown that also $\langle F/g \rangle \supseteq \mathscr{I} : \langle g \rangle$ holds, so that $\langle F/g \rangle = \mathscr{I} : \langle g \rangle$ or $F/g$, respectively, is a basis of $\mathscr{I} : \langle g \rangle$. □

### 2.1.6 Polynomial vs. Laurent ideals

In principle[10] things are very simple: any monomial is a unit in $\Lambda$ and this enables us to transform any Laurent ideal into a polynomial ideal. Nevertheless, we have to be careful as the simple example

$$f(x,y) = x\,y^{-1} - 1 = y^{-1}\left(x - y\right)$$

shows that coincides, up to units, with $\widetilde{f}(x,y) = x - y$. This looks innocent, but the "equivalent" polynomial $\widetilde{f}$ has a zero at $x = y = 0$, which is forbidden for Laurent polynomials. This discrepancy in varieties is a clear sign that we have to be more careful and find a more delicate approach.

[9] And not arbitrary rational functions.

[10] Principles are always dangerous, the only more dangerous thing is to insist in them.

**Definition 2.1.34.** The POLYNOMIAL PART $P(\mathscr{J})$ of a Laurent ideal $\mathscr{J} \subseteq \Lambda$ is defined as $P(\mathscr{J}) := \mathscr{J} \cap \Pi$.

**Exercise 2.1.3** Show that $P(\mathscr{I})$ is a *polynomial* ideal. ♢

Polynomial parts of Laurent ideals have a very special structure.

**Proposition 2.1.35.** *A polynomial ideal $\mathscr{I} \subseteq \Pi$ is of the form $\mathscr{I} = P(\mathscr{J})$ for a Laurent ideal $\mathscr{J}$ if and only if for $f \in \Pi$ and $1 \le j \le s$ we have that*

$$x_j f(x) \in \mathscr{I} \qquad \Rightarrow \qquad f \in \mathscr{I}. \tag{2.1.29}$$

**Proof:** Set $\mathscr{I} := P(\mathscr{J}) \subset \mathscr{J}$. Since $x_j^{-1} \in \Lambda$, $j = 1, \ldots, s$, it follows for $f \in \mathscr{I}$ that

$$x_j f(x) \in \mathscr{I} \subset \mathscr{J} \quad \Rightarrow \quad f(x) = x_j^{-1}\big(x_j f(x)\big) \in \mathscr{I} \cap \Pi = P(\mathscr{J}) = \mathscr{I}.$$

Hence, (2.1.29) is necessary for polynomial parts of Laurent ideals.

Conversely, let $\mathscr{I}$ be a polynomial ideal satisfying (2.1.29), $F \subset \mathscr{I}$ a basis of $\mathscr{I}$ and $\mathscr{J} := \langle F \rangle_\Lambda$, yielding

$$\mathscr{I} \subseteq \mathscr{J} \cap \Pi = P(\mathscr{J}).$$

If $P(\mathscr{J}) \neq \mathscr{I}$, there exist Laurent polynomials $g_f \in \Lambda$, $f \in F$, with the property that

$$g = \sum_{f \in F} g_f f \in P(\mathscr{J}) \setminus \mathscr{I}.$$

Choosing a monomial $m \in \Pi$ such that $m g_f \in \Pi$, $f \in F$, we get that

$$m g = \sum_{f \in F} \big(m g_f\big) f \in \langle F \rangle_\Pi = \mathscr{I},$$

and a repeated application of (2.1.29), namely, that for $\alpha \in \mathbb{Z}^s$ we have

$$x^\alpha f(x) \in \mathscr{I} \qquad \Rightarrow \qquad f \in \mathscr{F}$$

allows us to conclude from $m g \in \mathscr{I}$ the contradiction $g \in \mathscr{I}$. Therefore $\mathscr{I} = P(\mathscr{J})$. □

**Remark 2.1.36.** Proposition 2.1.35 tells us how to compute the polynomial parts of a Laurent ideal $\mathscr{J}$, again by means of completion: we start with a basis of $\mathscr{J}$, transform that into a basis consisting of polynomials by multiplying with proper monomials, and check if the polynomial ideal $\mathscr{I}$ generated by these polynomials satisfies (2.1.29), which we can rewrite

$$\mathscr{I} : \langle z_j \rangle = \mathscr{I}, \qquad j = 1, \ldots, n, \tag{2.1.30}$$

If *strict* inclusion holds in (2.1.30), we extend $\mathscr{I}$ into a basis of $\mathscr{I} : \langle z_j \rangle$, otherwise we have constructed a basis for the polynomial part $P(\mathscr{J})$.

**Example 2.1.37.** In the context of $\sqrt{3}$ *subdivision* [Kobbelt, 2000] one is interested in the ideal $\mathscr{J} = \langle x y^{-2} - 1, x^2 y^{-1} - 1 \rangle$ whose associated polynomial basis elements $x - y^2$ and $x^2 - y$ have the common but SPURIOUS ZERO $x = y = 0$. This can also be recognized in the polynomial part that is of the form

$$P(\mathscr{J}) = \langle y^2 - x, x^2 - y, x y - 1 \rangle = \langle y^2 - x, x^2 - y \rangle + \langle x y - 1 \rangle,$$

where the additional polynomial takes care of the unwanted zero.

**Proposition 2.1.38** (Laurent ideals)**.**

1. *A zero dimensional polynomial ideal $\mathscr{I}$ is the polynomial part $P(\mathscr{J})$ of a Laurent ideal $\mathscr{J}$ if and only if*
$$f(z) \neq 0, \qquad z \in \mathbb{C}_\times^s, \quad f \in \mathscr{I}. \tag{2.1.31}$$
2. *For Laurent ideals $\mathscr{J}$, $\mathscr{J}'$ one has*
$$P\big(\mathscr{J} : \mathscr{J}'\big) = P\big(\mathscr{J}\big) : P\big(\mathscr{J}'\big) \tag{2.1.32}$$
*and*
$$P\Big(\mathscr{J}^k\Big) = P\big(\mathscr{J}\big)^k, \qquad k \in \mathbb{N}. \tag{2.1.33}$$

**Proof:** Being zero dimensional, $\mathscr{I}$ has a finite associated variety $\mathscr{X} = V(\mathscr{I})$ and thus a primary decomposition
$$\mathscr{I} = \bigcap_{x \in \mathscr{X}} \langle z - x \rangle^{k_x}, \qquad k_x \in \mathbb{N}, \quad x \in \mathscr{X}.$$
Because of (2.1.30), $\mathscr{I} = P(\mathscr{J})$ for some Laurent ideal $\mathscr{J}$ iff
$$\langle z - x \rangle^{k_x} : \big\langle z_j \big\rangle = \langle z - x \rangle^{k_x}, \qquad j = 1, \ldots, n,\ x \in X,$$
which is in turn equivalent to
$$z_k \notin \sqrt{\langle z - x \rangle^{k_x}} = \langle z - x \rangle \qquad \Rightarrow \qquad x_k \neq 0,$$
which yields the first claim 1).

To prove (2.1.32) we choose $f \in P(\mathscr{J}) : P\big(\mathscr{J}'\big)$, this is, $f\, P\big(\mathscr{J}'\big) \subset P\big(\mathscr{J}\big)$ and therefore
$$\big\langle f \big\rangle_\Lambda \mathscr{J}' = \{ g \underbrace{(f h)}_{\in \mathscr{J}} : g \in \Lambda,\ h \in \mathscr{J}' \} \subset \mathscr{J}.$$
Consequently,
$$f \in \big(\mathscr{J} : \mathscr{J}'\big) \cap \Pi = P\big(\mathscr{J} : \mathscr{J}'\big)$$
which means that $P(\mathscr{J}) : P\big(\mathscr{J}'\big) \subseteq P\big(\mathscr{J} : \mathscr{J}'\big)$. Conversely, one has for any $f \in P\big(\mathscr{J} : \mathscr{J}'\big)$ that
$$\Pi \supset f\, P\big(\mathscr{J}'\big) \subset f\, \mathscr{J}' \subset \mathscr{J} \quad \Rightarrow \quad f\, P\big(\mathscr{J}'\big) \subset P\big(\mathscr{J}\big)$$
and therefore also $P(\mathscr{J}) : P\big(\mathscr{J}'\big) \supseteq P\big(\mathscr{J} : \mathscr{J}'\big)$.

Since, according to (2.1.19) $\mathscr{J}^k \subseteq \mathscr{J}$ holds for any Laurent ideal $\mathscr{J} \subseteq \Lambda$ and any $k \in \mathbb{N}$, we have $V\big(\mathscr{J}^k\big) \supseteq V(\mathscr{J})$. On the other hand, $f \in \mathscr{J}$ implies $f^k \in \mathscr{J}^k$ and $f^k(x) = 0$ also yields $f(x) = 0$, i.e. $x \in V(\mathscr{J})$, so that $V\big(\mathscr{J}^k\big) \supseteq V(\mathscr{J})$. Hence
$$V\Big(\mathscr{J}^k\Big) = V(\mathscr{J}) \subset \mathbb{C}_\times^n,$$
because of 1). The converse direction of 1) also implies that $P\big(\mathscr{J}^k\big)$ is the polynomial part of a Laurent ideal, due to which, according to Proposition 2.1.35,
$$P\Big(\mathscr{J}^k\Big) : \big\langle z_j \big\rangle = P\Big(\mathscr{J}^k\Big), \qquad j = 1, \ldots, n.$$
If $F$ is a basis of $P(\mathscr{J})$ and therefore also of $\mathscr{J}$, then the polynomials
$$\left\{ \prod_{f \in F} f^{k_f} \ : \ \sum_{f \in F} k_f = k \right\} \subset P\big(\mathscr{J}\big)^k$$
is also a basis of $\mathscr{J}^k$ and by Proposition 2.1.35 this yields (2.1.33). □

## 2.2 Degree: graded rings and polynomial degree

In one variable the notion of the degree of a polynomial was clear from intuition and has never been questioned: it is simple the maximal index of nonzero coefficients or, which is the same, the largest exponent appearing in the monomial representation. If we want to derive an analogy in several variables, the direct analogy would force us to order multiindices before the even more general question pops up: what *type* of object should a degree be? A minimal condition for a degree, and inspection of proofs shows that this is mostly what we use is that

$$\deg(f+g) \leq \max\{\deg f, \deg g\}, \qquad \deg\big(f\cdot g\big) = \deg f + \deg g, \qquad f,g \in \Pi. \tag{2.2.1}$$

Abstractly spoken, the concept of a degree connects the *multiplicative* structure of polynomials with the *additive* structure of $\mathbb{N}_0$, at least in the case of univariate polynomials. Which implies that degrees should at least be something for which an addition is well-defined. Such structures can be defined formally.

**Definition 2.2.1.** A MONOID $\Gamma$ is a commutative additive semigroup, i.e., closed under addition, with a neutral element 0.

**Example 2.2.2.** The sets $\mathbb{N}_0^s$ and $\mathbb{Z}^s$, $s \geq 1$, are monoids while $2\mathbb{N}$ is a semigroup but no monoid - it lacks the neutral element.

The definition of a graded ring, from which we will derive the notion of degree afterwards, consist precisely of impose the structure of a monoid on a ring in a consistent fashion. This will still mean that different monoids can lead to different notions of degree for the same ring $R$, even the same monoid can lead to different structures. This will give us a liberty of choice that we can use, for example, to derive numerically robust methods.

**Definition 2.2.3** (Graded ring & grading monoid)**.** A commutative ring $R$ with unit is called GRADED RING[11] if there exists a GRADING MONOID $\Gamma$ such that

1. $R$ has the direct sum decomposition

$$R = \bigoplus_{\gamma\in\Gamma} R_\gamma \tag{2.2.2}$$

   into additive subgroups $R_\gamma \subseteq R$, $\gamma \in \Gamma$, of $R$. Direct sum means that $R_\gamma \cap R_{\gamma'} = \{0\}$, $\gamma \neq \gamma'$, and that the representation

$$r = \sum_{\gamma\in\Gamma} r_\gamma, \qquad r_\gamma \in R_\gamma,$$

   is *unique*.

2. the summands have the property

$$R_\gamma \cdot R_{\gamma'} \subseteq R_{\gamma+\gamma'}, \qquad \gamma,\gamma' \in \Gamma. \tag{2.2.3}$$

We call any element of the summands $R_\gamma$, $\gamma \in \Gamma$, a HOMOGENEOUS ELEMENT of $R$ and write

$$R^0 = \bigcup_{\gamma\in\Gamma} R_\gamma \tag{2.2.4}$$

for the set of all homogeneous elements.

[11] Therefore, whenever we speak of a grading and a graded ring, *commutativity* and the *existence of a unit* are included.

**Remark 2.2.4.** The transfer of additive to multiplicative structures is, of course, property 2), more precisely (2.2.3).

**Example 2.2.5** (Gradings of $\mathbb{K}[x]$)**.** One can come up with various gradings for multivariate polynomials:

1. The TOTAL DEGREE is obtained with $\Gamma = \mathbb{N}_0$ and
$$\Pi_k := \operatorname{span}_{\mathbb{K}} \{x^\alpha : |\alpha| = k\}.$$
2. GRADING BY MONOMIALS uses $\Gamma = \mathbb{N}_0^n$ and
$$\Pi_\alpha = \operatorname{span}_{\mathbb{K}} \{x^\alpha\}.$$
3. For $s = 1$ the gradings in 1) and 2) coincide.
4. A more general grading can be obtained as follows: let $v_j \in \mathbb{K}^s$, $j = 1,\dots,s$, be linearly independent vectors and define the linear polynomials $\ell_j(x) = v_j^T x$, $x \in \mathbb{K}^s$, $j = 1,\dots,s$. Defining $\ell^\alpha := \ell_1^{\alpha_1} \cdots \ell_n^{\alpha_n}$, the homogeneous spaces
$$\Pi_k = \operatorname{span}_{\mathbb{K}} \{\ell^\alpha : |\alpha| = k\}, \qquad \Pi_\alpha = \operatorname{span}_{\mathbb{K}} \{\ell^\alpha\}, \qquad k \in \mathbb{N}_0, \quad \alpha \in \mathbb{N}_0^s,$$
both induce a grading that turns $\mathbb{K}[x]$ into a graded ring.

Having generalized the notion of a term into homogeneous elements, we need a largest homogeneous element to extend the concept of degree. This forces us to order the monoid.

**Definition 2.2.6** (Well ordering)**.** An ordering "$<$" on a monoid $\Gamma$ is called WELL ORDERING if

1. it is a TOTAL ORDER, that is
$$\gamma, \gamma' \in \Gamma \quad \gamma \neq \gamma' \qquad \Rightarrow \qquad \gamma < \gamma' \quad \text{or} \quad \gamma' < \gamma.$$
2. it is COMPATIBLE with the semigroup operation "$+$", that is,
$$\gamma < \gamma' \qquad \Rightarrow \qquad \gamma + \eta < \gamma' + \eta, \qquad \eta \in \Gamma.$$
3. each STRICTLY DESCENDING sequence $\gamma_1 > \gamma_2 > \cdots$ of monoid elements is finite.

**Remark 2.2.7.** Property 3) connects fundamentally to polynomial ideals where we have the so called ASCENDING CHAIN CONDITION: any strictly increasing sequence of ideals $\mathscr{I}_1 \subset \mathscr{I}_2 \subset \cdots$ has to be finite. Indeed, this is what makes polynomials a NOETHERIAN RING. For more details see once more [Cox et al., 1996][12].

It is easy to see that for any well ordering we know the minimal element in advance.

**Lemma 2.2.8.** *If "$<$" is a well ordering on the monoid $\Gamma$, then $0 < \gamma$ for all $\gamma \in \Gamma \setminus \{0\}$.*

**Proof:** Supposing that there exists $\gamma < 0$ we get
$$\gamma = \gamma + 0 > \gamma + \gamma =: 2\gamma = 2\gamma + 0 > 2\gamma + \gamma = 3\gamma > \cdots$$
and thus the strictly decreasing sequence $k\gamma$, $k \in \mathbb{N}$, is infinite, in contradiction to Definition 2.2.6, 3). □

[12] The reader may get the idea that this a good book and worthwhile to read. I am not objecting.

**Definition 2.2.9** (Term order)**.** A grading on $\mathbb{K}[x]$ is called a TERM ORDER if $\Gamma = \mathbb{N}_0^s$ and $\Pi_\alpha = \operatorname{span}\{x^\alpha\}$, $\alpha \in \mathbb{N}_0^n$.

**Remark 2.2.10** (Well orderings)**.**

1. The only well ordering on $\mathbb{N}_0$ is the canonical one since $0 < 1$ already implies $k < k+1 < k+2 < \cdots$ for any $k \in \mathbb{N}_0$.
2. There is no well ordering on $\mathbb{Z}$ as Lemma 2.2.8 enforces $k > 0$ and $-k > 0$ for any $k \neq 0$, leading to the contradiction $0 < k + (-k) = 0$ due to the compatibility property 2).
3. On $\mathbb{N}_0^s$ there exists a multitude of well orderings from which one can choose the most appropriate for a specific application. The classics are:
   a) LEXICOGRAPHICAL term order ("lex"): for $\alpha \neq \beta \in \mathbb{N}_0^n$ we set
   $$\alpha \prec_l \beta \qquad \Leftrightarrow \qquad \alpha_j = \beta_j, \quad j = 1, \ldots, k-1, \quad \alpha_k < \beta_k.$$
   b) GRADED LEXICOGRAPHICAL term order ("gradlex"): for $\alpha \neq \beta \in \mathbb{N}_0^n$ we define
   $$\alpha \prec_g \beta \qquad \Leftrightarrow \qquad |\alpha| < |\beta| \quad \text{oder} \quad |\alpha| = |\beta|, \alpha \prec_l \beta.$$
   In the "gradlex" ordering the lexicographical ordering plays the role of "tie breaker" for multiindices of the same length.

There is an important interaction between grading and units.

**Lemma 2.2.11.** *If $R$ is a graded by a well ordered monoid, it units satisfy $R^\times \subseteq R_0$.*

**Proof:** Write $r \in R^\times$ and its inverse $s = r^{-1}$ with respect to the direct sum decomposition as

$$r = \sum_{\gamma \in \Gamma} r_\gamma \qquad \text{and} \qquad s = \sum_{\gamma \in \Gamma} s_\gamma,$$

i.e., with respect to its homogeneous components. Then we get for $\eta \in \Gamma$,

$$R_\eta = 1 \cdot R_\eta = (rs)\, R_\eta = \sum_{\gamma, \gamma' \in \Gamma} \underbrace{r_\gamma s_{\gamma'} R_\eta}_{\in R_{\gamma+\gamma'+\eta}},$$

and since $\gamma + \gamma' > 0$ if $\gamma, \gamma' \neq 0$, the uniqueness of the direct sum decomposition implies that $r_\gamma = s_\gamma = 0$ for $\gamma \in \Gamma \setminus \{0\}$. □

**Corollary 2.2.12.** *The only grading for Laurent polynomials is the* TRIVIAL GRADING *$\Lambda = R_0$ and $R_\gamma = \{0\}$, $\gamma \in \Gamma \setminus \{0\}$.*

**Proof:** As a vector space, $\Lambda$ is generated by units that belong to $R_0$, so all elements of $\Lambda$ belong to $R_0$. □

**Exercise 2.2.1** Show that the trivial grading is a grading (easy). ♢

Sometimes gradings will be interesting where $R_0$ is as small as possible, somehow the counterpiece to the trivial grading.

**Definition 2.2.13** (Strict grading)**.** A grading is called a STRICT GRADING if $R_0 = R^\times$.

Returning to our concrete ring $\Pi = \mathbb{K}[x]$, we can now use the concept of the graded ring with a well ordered monoid to finally define a degree, just recalling that the degree was the *maximal* index of a nontrivial homogeneous component. And this we have for any well ordered grading monoid for $\Pi$.

**Definition 2.2.14** (Degree)**.** Let $\Gamma$ be a well ordered grading monoid for $\mathbb{K}[x]$. For the[13]

$$\Pi \ni f = \sum_{\gamma\in\Gamma} f_\gamma \qquad f_\gamma \in \Pi_\gamma, \tag{2.2.5}$$

we define

1. the ($\Gamma$-)DEGREE of $f$ as
$$\delta_\Gamma(f) := \max\{\gamma \in \Gamma \,:\, f_\gamma \neq 0\} \quad \in \Gamma. \tag{2.2.6}$$
2. the ($\Gamma$-)LEADING PART of $f$ as
$$\lambda_\Gamma(f) := f_{\delta_\Gamma(f)} \quad \in \Pi^0. \tag{2.2.7}$$

**Exercise 2.2.2** Show that the decomposition (2.2.5) contains only finitely many nonzero terms. ♢

**Exercise 2.2.3** Show that $\delta_\Gamma(f\cdot g) = \delta_\Gamma(f) + \delta_\Gamma(g)$ and $\lambda_\Gamma(f\cdot g) = \lambda_\Gamma(f)\cdot\lambda_\Gamma(g)$. ♢

**Remark 2.2.15.**

1. We can also consider the degree and the leading part as mappings for $\Pi$ to $\Gamma$ and $\Pi^0$, respectively.
2. In the case of a term order, the leading part is a term, i.e., a multiple of a monomial, and thus usually called the LEADING TERM of the polynomial. In the case of the total degree the leading part is a HOMOGENEOUS POLYNOMIAL or FORM and therefore called the LEADING FORM.
3. Strictly speaking, degree and leading part depend of the monoid $\Gamma$ *and* the well ordering "$<$" which *together* form the grading monoid. For example, if we consider the polynomial $f(x,y) = 2x^2y^2 + 3x^3$ with the "lex" grading, then $\delta(f) = (3,0)$ and $\lambda(f) = 3x^3$, while "gradlex" gives $\delta(f) = (2,2)$ and $\lambda(f) = 2x^2y^2$.

To get acquainted to this somewhat lesser known concept, let us consider some more examples.

**Example 2.2.16** (Weighted total degree)**.** We choose $0 \neq \omega \in \mathbb{N}_0^s$, $\Gamma = \mathbb{N}_0$ as monoid[14] and

$$\Pi_k = \operatorname{span}_{\mathbb{K}}\{x^\alpha \,:\, \omega^T\alpha = k\}, \qquad k \in \mathbb{N}_0.$$

If $\omega = (1,\dots,1)$, we rediscover the total degree. The associated grading is also called the H-GRADING where "H" stands for "homogeneous".

It is *not* forbidden here that $\omega_j = 0$ for one or several values of $j$. If, for example, $\omega_2 = \cdots = \omega_n = 0$, the we consider the polynomial only as a polynomial in $x_1$ and take the usual degree of that one. In particular, $\Pi_0$ consists of all linear combinations if monomials of the form $x_2^{\alpha_2}\cdots x_n^{\alpha_n}$ and is therefore of infinite dimension. And the grading is not strict any more.

[13] It is unique by definition, see Definition 2.2.3, 1).

[14] There is only one well ordering for this one ...

**Example 2.2.17** (Matrix grading)**.** What we could do with vectors[15], can also be done with matrices, for example in the following setup: let $m \in \mathbb{N}$, $0 \neq M \in \mathbb{N}_0^{m\times s}$, let $\prec$ be a well ordering on $\mathbb{N}_0^m$. The we set

$$\Pi_\beta = \operatorname{span}_{\mathbb{K}} \left\{ x^\alpha \,:\, M\alpha = \beta \right\}, \qquad \beta \in \mathbb{N}_0^m.$$

Here it can happen that $\Pi_\beta = \{0\}$ for some values of $\beta$. A few particular cases are as follows:

1. The case $m = 1$ is the situation of Example 2.2.16.
2. For $M = 2I$ we get that $\Pi_\beta = \{0\}$ whenever $\beta_j \in 2\mathbb{N}+1$ for at least one $j \in \{1,\dots,s\}$.
3. Setting $m = s+1$ and choosing $\prec$ as the lex ordering on $\mathbb{N}_0^m$, the matrix

$$M = \begin{bmatrix} 1 & \dots & 1 \\ 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix}$$

   describes the gradlex ordering. Indeed, any term order can be reduced to the lexicographic one by means of matrix multiplication which allows us to parameterize term orders.
4. If $\prec_m$ is a well ordering on $\mathbb{N}_0^s$ and $M \in \mathbb{N}_0^{m\times s}$, then the ordering $\prec_s$ on $\mathbb{N}_0^s$, defined by

$$\alpha \prec_n \beta \qquad \Longleftrightarrow \qquad M\alpha \prec_m M\beta, \qquad \alpha,\beta \in \mathbb{N}_0^n,$$

   is a well ordering if and only if $\ker_{\mathbb{Z}^s} M = \left\{ \alpha \in \mathbb{Z}^s \,:\, M\alpha = 0 \right\} = \{0\}$; this requirement is needed to able to compare two different elements, hence of a total order.

**Example 2.2.18.** It is even possible to grade with $\mathbb{R}$, more precisely with a finitely generated submonoid of $\mathbb{R}$. To that end, let $\omega \in \mathbb{R}_+^s$ be a vector whose components are linearly independent over $\mathbb{Q}$, that is,

$$\left\{ q \in \mathbb{Q}^n \,:\, \omega^T q = 0 \right\} = \{0\},$$

for example $\omega = \left(1, \sqrt{2}, \pi\right)^T$. Then

$$\alpha \prec \beta \qquad \Longleftrightarrow \qquad \underbrace{\omega^T\alpha}_{\in\mathbb{R}} < \underbrace{\omega^T\beta}_{\in\mathbb{R}}$$

is even a term order

**Definition 2.2.19.** A grading is called MONOMIAL GRADING if all homogeneous spaces $\Pi_\gamma$, $\gamma \in \Gamma$ are spanned by monomials as a $\mathbb{K}$ vector space.

## 2.3 Division with remainder: making the impossible possible

The simple idea behind efficient, computable bases, is to lift the division by remainder to the multivariate situation. In principle, this is impossible since $\mathbb{K}[x]$ is a euclidean ring only if $s = 1$, i.e., in the univariate case. Nevertheless, this should not prevent us from trying to naively extend the algorithm, then analyze the problems and find a way to overcome them. Since the approach severely relies on the notion of a degree, we will only consider *polynomial* ideals in this section.

[15] Which are in fact only tuples in Example 2.2.16.

### 2.3.1 A different perspective ...

The key to a successful *algorithmic* handling of ideal bases will be yet another division with remainder. This is not fully intuitive and does not suggest itself since for $s > 1$ the ring $\mathbb{K}[x]$ is **not** a euclidean ring any more. This forces us to change our point of view a little bit and consider division with remainder from a different perspective. As we have seen before, in $s = 1$ for any given $f \in \mathbb{K}[x]$ each polynomial $g \in \mathbb{K}[x]$ can be written as

$$g(x) = p(x)\, f(x) + r(x), \qquad \deg r < \deg f, \tag{2.3.1}$$

where uniqueness of the REMAINDER $r$ was a consequence of the degree requirement that is imposed in (2.3.1). Since this in turn was due to the fact that the degree is a euclidean function, cf. [Gathen and Gerhard, 1999] we have to consider things from a more general point of view if want to generalize the idea:

1. the polynomials $p(x)\, g(x)$ is an element of the principal ideal $\langle g \rangle$.
2. If we assume that $f(x) = f_n x^n + \cdots$ then $\deg r < \deg f$ is also equivalent to the fact that $r$ contains no term of the form $r_n x^n, r_{n+1} x^{n+1}, \ldots$, that is, no term that is a multiple of the leading term $\lambda(f)$ of $f$.

Divisibility by leading terms is now a concept that we can extend to the multivariate case as long as we work on terms only. Let us illustrate this idea by means of a (very) simple example[16].

**Example 2.3.1.** Let us fix $\alpha \in \mathbb{N}_0^n$ and consider division with remainder by the principal ideal $F = \{x^\alpha\}$. Since any polynomial

$$g(x) = \sum_{\beta \in \mathbb{N}_0^n} g_\beta\, x^\beta \quad \in \Pi$$

can be decomposed into

$$g(x) = \underbrace{\left( \sum_{\beta \in \alpha + \mathbb{N}_0^n} g_\beta\, x^{\beta - \alpha} \right) x^\alpha}_{\in \langle F \rangle} + \sum_{\beta \in \mathbb{N}_0^n \setminus (\alpha + \mathbb{N}_0^n)} g_\beta\, x^\beta, \tag{2.3.2}$$

hence an ideal and a divisible part, the monomial $f(x) = x^\alpha$ splits the support of $f$ into two parts: the terms divisible by $x^\alpha$ and those not divisible by the monomial. This decomposition is illustrated in Fig. 2.3.1.

To advance this idea, we need some further concepts that will be introduced in the next subsection.

### 2.3.2 Upper and lower sets and monomial ideals

It is never a bad a idea to begin by defining the objects one is going to study. In particular, as they will play a fundamental role later.

**Definition 2.3.2** (Upper/lower sets & monomial ideals)**.**

[16] As we will see soon, the example is even *too* simple.

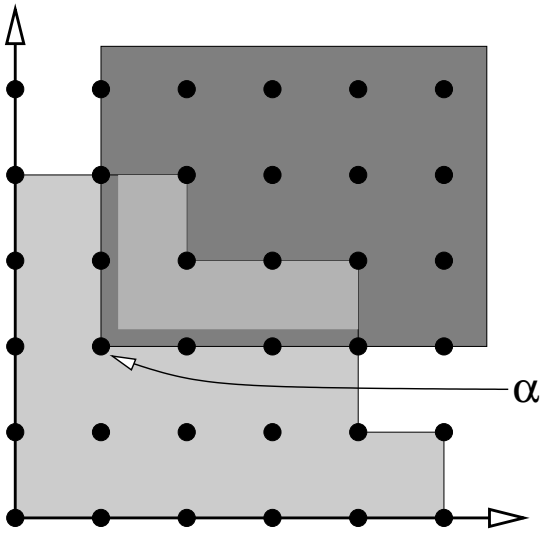



Figure 2.3.1: The exponents that belong to the cone $\alpha + \mathbb{N}_0^2$ spanned by $\alpha = (1,2)$ and those that do not belong to this cone. This yields the decomposition of the support of a polynomial (light gray region) into a "division" and a "remainder" part as in (2.3.2).

1. For a subset $A \subset \mathbb{N}_0^s$ we denote by $x^A = \{x^\alpha : \alpha \in A\}$ the set of all monomials indexed by $A$ and by $\Pi_A = \operatorname{span} x^A$ the vector space spanned by these monomials.

2. $A \subset \mathbb{N}_0^s$ is called a LOWER SET if

$$\alpha \in A \qquad \Rightarrow \qquad \{\beta : \beta \leq \alpha\} \subseteq A \tag{2.3.3}$$

   and it is called an UPPER SET if

$$\alpha \in A \qquad \Rightarrow \qquad \alpha + \mathbb{N}_0^s \subseteq A. \tag{2.3.4}$$

3. An ideal $\mathscr{I} \subset \Pi$ is called a MONOMIAL IDEAL if there exists some set $A \subseteq \mathbb{N}_0^s$ such that $\mathscr{I} = \langle x^A \rangle$.

**Remark 2.3.3** (Upper and lower sets)**.**

1. There exist various names for upper sets, for example ORDER CLOSED IDEAL.

2. Moreover, any upper set is clearly of infinite cardinality, for lower sets this can vary: $\mathbb{N}_0^s$ and $\{0\}$ are both (extreme) forms of nontrivial lower sets.

3. The principal lower and upper sets are of the form

$$L(\alpha) := \left\{\beta \in \mathbb{N}_0^s : \beta \leq \alpha\right\} \qquad \text{and} \qquad U(\alpha) := \alpha + \mathbb{N}_0^s, \tag{2.3.5}$$

   for some $\alpha \in \mathbb{N}_0^s$. $L(\alpha)$ is a higher dimensional cuboid, $U(\alpha)$ a cone, more precisely, a shifted octant.

**Lemma 2.3.4.** *Lower and upper sets are closed under union and intersection and the complement of an upper set is a lower set and vice versa.*

**Proof:** If $L, L'$ are lower sets and and $\alpha \in L \cap L'$, then $L(\alpha) \subseteq L$ and $L(\alpha) \subset L'$, hence $L(\alpha) \subseteq L \cap L'$; for $\alpha \in L \cup L'$ suppose that $\alpha \in L$, then $L(\alpha) \subseteq L \subseteq L \cup L'$. Literally the same proof, replacing $L$ by $U$ works for upper sets.

If $U$ is an upper set and $\alpha \notin U$ the assumption $L(\alpha) \cap U \neq \emptyset$ would result in the existence of some $\beta \in U$ with $\beta \leq \alpha$, hence $\alpha \in U(\beta) \subseteq U$, which is a contradiction. The same way, we get for a lower set $L$ and $\alpha \notin L$ that $\beta \in U(\alpha) \cap L$ would result in $\alpha \in L(\beta) \subseteq L$. □

**Remark 2.3.5.** The above argument about complements is based on the simple observation that

$$\alpha \in L(\beta) \qquad \Leftrightarrow \qquad \beta \in U(\alpha) \tag{2.3.6}$$

for any $\alpha, \beta \in \mathbb{N}_0^s$.

**Definition 2.3.6.** The BORDER of a lower set $L \subset \mathbb{N}_0^s$ is defined as[17]

$$\partial L := \left( \bigcup_{j=1}^{s} L + \epsilon_j \right) \setminus L. \tag{2.3.7}$$

**Theorem 2.3.7** (Upper sets)**.** *Let $U \subseteq \mathbb{N}_0^s$ be an upper set.*

1. *$U$ is generated by the border if its complement, that is,*

$$U = U(\partial L) = \bigcup_{\alpha \in \partial L} U(\alpha), \qquad L := \mathbb{N}_0^s \setminus U, \tag{2.3.8}$$

   *with the convention that $\partial \emptyset = \{0\}$.*

2. *Any upper set is finitely generated, i.e., there exists $A \subset \mathbb{N}_0^s$, $\#A < \infty$, such that $U = U(A)$.*

3. *Any upper set $U$ is minimally generated by the finite set*

$$G(U) := \{\alpha \in U : L(\alpha) \cap U = \{\alpha\}\}, \qquad \textit{i.e.,} \qquad U = U(G(U)). \tag{2.3.9}$$

**Proof:** For 1) we choose any $\alpha \in U$ and check whether $L(\alpha) \cap U \neq \{\alpha\}$. If yes, we replace $\alpha$ by one element of this set which we can do only finitely many times since we strictly decrease one column of $\alpha$ in each step. After finitely many steps we thus find an element $\beta$ such that $\alpha \in U(\beta)$ and $\beta$ cannot be reduced any further. This means that either $\beta = 0$ in which case $U = \mathbb{N}_0^s$ and the claim holds by convention, otherwise $\beta - \epsilon_j \notin U$, hence $\beta - \epsilon_j \in L$ even for all $j$ such that $\beta_j > 0$, so that indeed $\beta = \gamma + \epsilon_j$ for some $\gamma \in L$ and $j \in \{1, \ldots, s\}$.

For 2) we modify the proof of [Cox et al., 1996, §4, Theorem 5] to fit for subsets of $\mathbb{N}_0^s$. To that end, we perform by induction on $s$, where the case $s = 1$ is easy: any upper set of the form $\{k \in \mathbb{N} : k \geq m\}$ and $m = \min U$ is the generator[18] of the upper set. So suppose that the claim has been verified for some $s \geq 1$ and let $U \subseteq \mathbb{N}_0^{s+1}$ be an upper set in $\mathbb{N}_0^{s+1}$. Define

$$\mathbb{N}_0^s \ni U' := \left\{\alpha \in \mathbb{N}_0^s : (\alpha, k) \in U \text{ for some } k \in \mathbb{N}_0\right\}$$

as the projection of $U$ on $\mathbb{N}_0^s$. The set $U'$ is an upper set since $U$ is an upper set, hence, by induction, it is generated by a finite set $A'$, $U' = U(A')$. By definition of $U'$ there exist uniquely

[17] Note that we always use the SET DIFFERENCE in the sense that $A \setminus B = \{a \in A : a \notin B\}$ and do not require that the "removed" set $B$ is originally a subset of $A$.

[18] This reminds us a lot of principal ideals and that is no accident.

defined *minimal*[19] numbers $k_\alpha$, $\alpha \in A'$, such that $(\alpha, k_\alpha) \in U'$; set $k^* = \max\{k_\alpha : \alpha \in A'\}$. Next, we define the "slices"

$$U_k = \left\{\alpha \in \mathbb{N}_0^s : (\alpha, k) \in U\right\}, \qquad k = 0, \dots, k^* - 1,$$

which are upper sets again and therefore, by induction, there exist $A'_k \subset \mathbb{N}_0^s$, such that

$$U_k = U(A'_k), \qquad \#A'_k < \infty, \qquad k = 0, \dots, k^* - 1.$$

We now claim that

$$U = U(A), \qquad A := \left\{(\alpha, k_\alpha) : \alpha \in A'\right\} \cup \bigcup_{k=0}^{k^*-1} \left\{(\alpha, k) : \alpha \in A'_k\right\}. \tag{2.3.10}$$

Indeed, if $(\beta, k) \in U$, then $\beta \in U'$. If $k \geq k^*$, we use the fact that

$$\beta \in \bigcup_{\alpha \in A'} U(\alpha) \qquad \beta \in U(\alpha) \text{ for some } \alpha \in A'$$

to write $(\beta, k)$ as

$$(\beta, k) = (\alpha, k_\alpha) + \underbrace{\left(\beta - \alpha, k - k_\alpha\right)}_{\in \mathbb{N}_0^{s+1}} \in U\left((\alpha, k_\alpha)\right),$$

while for $k < k^*$ we use the fact that now $\beta \in U(A'_k)$, hence $\beta \in U(\alpha)$ for some $\alpha \in A'_k$ and thus

$$(\beta, k) = (\alpha, k) + (\beta - \alpha, 0) \in U\left((\alpha, k)\right),$$

which verifies (2.3.10) and thus completes the proof of part 2).

For 3) we first note that any generator of $U$ must generate $G(U)$ since $\alpha \in U(\beta)$ for some $\alpha, \beta \in U$ implies that $\beta \in L(\alpha)$, hence $G(U) \subseteq A$ for the generating set $A$ from 2). Hence, all we have to show is that $G(U)$ is finite and generates $U$. For any $\alpha \in A$ from the finite generating set $A$ we can find by the construction used to prove statement 1), a $\beta \in G(U)$ such that $\alpha \in U(\beta)$, hence there exists a *finite*[20] set $B \subseteq G(U)$ such that $A \subset U(G(U))$, hence $U = U(A) = U(B)$. This implies that $G(U) \subseteq B \subseteq G(U)$, hence $B = G(U)$ is the finite minimal generator for $U$. □

**Remark 2.3.8.** Statement 2) in Theorem 2.3.7 is known as DICKSON'S LEMMA. Among others, we will use it later to give a constructive proof of Hilbert's Basissatz, Theorem 2.1.29.

To apply Theorem 2.3.7 to (monomial) ideals, we need to make the connection between monomial ideals and upper sets. This is quite a direct one.

**Proposition 2.3.9.** *A monomial $x^\beta$ lies in the monomial ideal $\langle x^A \rangle$ if and only if $\beta \in U(A)$.*

**Proof:** The direction "⇐" follows from the definition as $\beta = \alpha + \gamma$ for some $\alpha \in A$ and $\gamma \in \mathbb{N}_0^s$ yields

$$x^\beta = x^{\alpha+\gamma} = x^\alpha\, x^\gamma \in \langle x^A \rangle.$$

For the converse direction, we note that $x^\beta \in \langle x^A \rangle$ implies that

$$x^\beta = \sum_{\alpha \in A} p_\alpha(x)\, x^\alpha$$

and since $\operatorname{supp} p(\cdot)^\alpha \subseteq U(\alpha)$, we get that

$$x^\beta \in \bigcup_{\alpha \in A} U(\alpha) = U(A)$$

which proves "⇒". □

[19] Since $U$ is an upper set, $(\alpha, k) \in U$ implies that $(\alpha, k+1) \in U$.
[20] At most one $\beta$ for each element of the finite set $A$

**Corollary 2.3.10.** *Any monomial ideal is finitely generated.*

**Proof:** The monomials contained in $\langle x^A \rangle$ are exactly $x^{U(A)}$ and even if $A$ is infinite, there exists a finite $A'$ such that $U(A) = U(A')$, then $\langle x^A \rangle = \left\langle x^{A'} \right\rangle$. □

### 2.3.3 Division with remainder: a naive monomial algorithm

After that little bit of theory on upper and lower sets, we return to naively performing division with remainder. A few names for that purpose that fix the intuition from the beginning of this section.

**Definition 2.3.11.** Let $F \subset \Pi$, $\#F < \infty$, and choose the grading monoid $\Gamma = \mathbb{N}_0^n$ so that all homogeneous spaces consist of single monomials.

1. By
$$\lambda(F) = \lambda_\Gamma(F) = \{\lambda_\Gamma(f) : f \in F\}$$
we denote the leading terms appearing in $F$.

2. We say that the finite set $F$ or the ideal $\langle F \rangle$ DIVIDES $g \in \Pi$ with REMAINDER $r$, if there are polynomials $g_f \in \Pi$ such that
$$g = \sum_{f \in F} g_f f + r, \tag{2.3.11}$$
and no element of $\lambda(F)$ divides any homogeneous component of $r$, i.e., any term in $r$.

3. We call a representation (2.3.11) of $g$ a G-REPRESENTATION if it also satisfies the DEGREE CONSTRAINT
$$\delta_\Gamma(g) \geq \delta_\Gamma\left(g_f f\right), \qquad \delta_\Gamma(g) \geq \delta_\Gamma(r). \tag{2.3.12}$$

**Remark 2.3.12.** One intuition for a G-representation is that it is *efficient* since it does not contain unnecessary redundant terms. Imaging one summand on the right hand side of (2.3.11) has a larger degree than $g$, then there must be another summand compensating that and the excessive degree was simply unnecessary and should have been omitted from the beginning.

The "G" in the G-representation relates, no surprise, to Gröbner bases. And even if we do not know yet whether something like that exists, we can already define this most important concept.

**Definition 2.3.13.** A finite set $F \subset \mathscr{I}$ is called a GRÖBNER BASIS for the ideal $\mathscr{I}$ if any $g \in \mathscr{I}$ has a G-representation with respect to $F$, i.e.,
$$g = \sum_{f \in F} g_f f, \qquad \delta(g) \geq \delta(g_f f). \tag{2.3.13}$$

Note that the representation in (2.3.11) is not really well-defined.

**Remark 2.3.14.** Neither the "coefficients" $g_f$ nor the remainder $r$ in (2.3.11) are unique. For example,
$$F = \{xy - 2, x^2 + 2y - 1\} =: \{f_1, f_2\} \quad \subset \mathbb{K}[x, y],$$

and $g(x) = x^2 y$ admit the two representations

$$\begin{array}{rcrlcrlcl} g(x,y) & = & g_1(x,y) & f_1(x,y) & + & g_2(x,y) & f_2(x) & + & r(x,y) \\ x^2 y & = & x & \left(xy-2\right) & + & 0 & \left(x^2+2y-1\right) & + & 2x \\ & = & 0 & \left(xy-2\right) & + & y & \left(x^2+2y-1\right) & + & y-2y^2. \end{array}$$

For the lex order they both are even G-representation so that we need not hope to achieve uniqueness by simply bounding the degree.

The first step is an algorithm that naively extends the idea of division with remainder to term orders. The only difference is that in Algorithm 2.3.1 we do not only divide by a single polynomial but by a finite set that defines an ideal. This is reasonable since we are not working in an principal ideal ring any more.

**Algorithm 2.3.1** DIVISION WITH REMAINDER; $g \in \Pi$, $F \subset \Pi$, term order $\Gamma$

1: $r \leftarrow 0, \qquad g_f \leftarrow 0,\ f \in F.$
2: **while** $g \neq 0$ **do**
3: **if** Exists $f \in F$ such that $\lambda(f) | \lambda(g)$ **then**
4:

$$g_f \leftarrow g_f + \frac{\lambda(g)}{\lambda(f)}, \qquad g \leftarrow g - \frac{\lambda(g)}{\lambda(f)} f \tag{2.3.14}$$

5: **else**
6:

$$r \leftarrow r + \lambda(g), \qquad g \leftarrow g - \lambda(g) \tag{2.3.15}$$

7: **end if**
8: **end while**
9: Result:

$$g = \sum_{f \in F} g_f f + r, \qquad \delta(g) \geq \delta(g_f f), \delta(r). \tag{2.3.16}$$

**Remark 2.3.15.** We can rephrase the crucial steps in Algorithm 2.3.1 also in terms of our upper and lower set terminology. In a term order the check whether $\lambda(f)$ divides $\lambda(g)$ for some $f$ is exactly the question whether $\delta(g) \in U\left(\delta(f)\right)$ for some $f \in F$, hence we check whether

$$\delta(g) \in U\left(\delta(F)\right). \tag{2.3.17}$$

If this is the case, then we reduce by means of the function $f$. Note that

1. this adds multiples of lower degree components of $f$ to $g$ as soon as $f$ is not only a monomial
2. this modification is no more unique as soon as there are several polynomials $f$ with $\delta(g) \in U\left(\delta(f)\right)$, and this can and will effect further steps of the algorithm.

If, on the other hand (2.3.17) is not true, we move the leading term directly into the remainder and thus have that

$$\operatorname{supp} r \subseteq \mathbb{N}_0^s \setminus U\left(\delta(F)\right) \tag{2.3.18}$$

is automatically a lower set. This simple observation will become relevant later.

Of course, we have to verify that the algorithm keeps what it promises.

**Proposition 2.3.16.** *Algorithm 2.3.1 terminates after finitely many steps and computes a G-representation* (2.3.16).

**Proof:** We denote by $g_j$ the value of $g$ in the $j$th step of Algorithm 2.3.1, initialized by $g_0 := g$ and with the iteration $g_{j+1} = g - \frac{\lambda(g_j)}{\lambda(f)} f$ and $g_{j+1} = g_j - \lambda(g_j)$ in (2.3.14) and (2.3.15), respectively.

Since in each step of the algorithm the leading of $g_j$ is eliminated, we immediately have that

$$\delta\left(g_{j+1}\right) < \delta\left(g_j\right), \qquad j = 0,1,2,\ldots$$

so that after finitely many steps we have to arrive at the zero polynomial since any grading is based on a well ordering. Therefore the algorithm terminates.

The more important property, namely (2.3.16) is proved by verifying the invariant

$$g = g_j + \sum_{f\in F} g_f f + r, \qquad j = 0,1,2,\ldots \tag{2.3.19}$$

which is trivially true for $j = 0$ by the way how $g_f$ and $r$ are initialized. Proceeding inductively, we observe that if $\delta(g_j) \in U(\delta(F))$, then, denoting the updated $g_f$ by $\tilde{g}_f$,

$$\begin{aligned} g_{j+1} + \sum_{f\in F} \tilde{g}_f f + r &= \left(g_j - \frac{\lambda\left(g_j\right)}{\lambda(f)} f\right) + \sum_{f'\in F\setminus\{f\}} g_{f'} f' + \left(g_f + \frac{\lambda\left(g_j\right)}{\lambda(f)}\right) f \\ &= g_j + \sum_{f\in F} g_f f + r, \end{aligned}$$

while for $\delta(g_j) \notin U(\delta(F))$

$$g_{j+1} + \sum_{f\in F} \tilde{g}_f f + \tilde{r} = \left(g_j - \lambda\left(g_j\right)\right) + \sum_{f\in F} \tilde{g}_f f + \left(r + \lambda\left(g_j\right)\right) = g_j + \sum_{f\in F} g_f f + r,$$

which advances (2.3.19) from $j$ to $j+1$. The degree constraints in (2.3.16), on the other hand, are enforced by the design of the algorithm. □

**Example 2.3.17.** To illustrate how the algorithm works, we consider the polynomials $F = \left\{xy-2, x^2+2y-1\right\} = \{f_1, f_2\}$ and the gradlex order with $x > y$ from Remark 2.3.14.

1. For $g(x,y) = x^3+y^3+1$ we get[21]

| $j$ | $g$ | $\lambda(g)$ | $g_{f_1}$ | $g_{f_2}$ | $r$ |
|---|---|---|---|---|---|
| 0 | $x^3+y^3+1$ | $x^3$ | 0 | 0 | 0 |
| 1 | $y^3-2xy+x+1$ | $y^3$ | 0 | $\boxed{x}$ | 0 |
| 2 | $-2xy+x+1$ | $-2xy$ | 0 | $x$ | $\boxed{y^3}$ |
| 3 | $x+5$ | $x$ | $\boxed{-2}$ | $x$ | $y^3$ |
| 4 | 5 | 5 | $-2$ | $x$ | $\boxed{y^3+x}$ |
| 5 | 0 | 0 | $-2$ | $x$ | $\boxed{y^3+x+5}$ |

   and the G-representation

$$g = -2\left(xy-2\right) + x\left(x^2+2y-1\right) + y^3 + x + 5$$

   of $g$ with respect to $F$.

[21] We always mark the objects that are updated.

2. For $g(x,y) = x^2y$ we get

| $j$ | $g$ | $\lambda(g)$ | $g_{f_1}$ | $g_{f_2}$ | $r$ |
|---|---|---|---|---|---|
| 0 | $x^2y$ | $x^2y$ | 0 | 0 | 0 |
| 1 | $2x$ | $2x$ | $\boxed{x}$ | 0 | 0 |
| 2 | 0 | 0 | $x$ | 0 | $\boxed{2x}$ |

or

| $j$ | $g$ | $\lambda(g)$ | $g_{f_1}$ | $g_{f_2}$ | $r$ |
|---|---|---|---|---|---|
| 0 | $x^2y$ | $x^2y$ | 0 | 0 | 0 |
| 1 | $-2y^2+y$ | $-2y^2$ | 0 | $\boxed{y}$ | 0 |
| 2 | $y$ | $y$ | 0 | $y$ | $\boxed{-2y^2}$ |
| 3 | 0 | 0 | 0 | $y$ | $\boxed{-2y^2-y}$ |

depending on whether we use $f_1$ or $f_2$ in the first step.

This second example above already helps us to see the problem of our naive extension: neither the G-representation nor the remainder and unique and thus well-defined in general. But we also can identify the source of this ambiguity. If in some step of Algorithm 2.3.1 there are *several* $f \in F$ such that $\delta(g) \in U(\delta(f))$, then different choices lead to different values of $g$ with which the algorithm proceeds and may come to different results. This is something that cannot happen in a principal ideal ring.

Of course, we could make the algorithm unique and the results well-defined by *numbering* the elements of $F$ as $f_1, \ldots, f_n$ and always choose the *first* $f$ from the list whose leading term divides the leading term of $g$. While this is possible and indeed leads to a well-defined behavior of the algorithm, it is not satisfactory at all as now the algorithm depends on the order of $F$ which is usually irrelevant and totally arbitrary.

What is so important about unique representation and, in particular, unique remainders. One reason is that unique remainders help us to solve a fundamental problem in computational ideal theory.

**Definition 2.3.18.** The IDEAL MEMBERSHIP PROBLEM consist of determining for given $g \in \Pi$ and $F \subset \Pi$ whether $g \in \langle F \rangle$

Indeed, whenever a G-representation ends up with $r = 0$, then $g$ can be represented with respect to $F$ and therefore belongs to $\langle F \rangle$. And if we had a *unique* remainder, the $r = 0$ would be the only possible remainder for any ideal element and the ideal membership is solved by simply computing a division with remainder. The good news is that there are bases which have this property and that we already know them by hearsay from (2.3.13) in Definition 2.3.11. The following result will not be proved now, as it will follow from a slight more general one that we will state and prove a little bit later.

**Theorem 2.3.19.** *If $G$ is a Gröbner basis for $\langle G \rangle$ in $f \in \Pi$ admits the G-representations*

$$f = \sum_{g \in G} f_g\, g + r = \sum_{g \in G} f'_g\, g + r',$$

*then $r = r'$.*

Based on Theorem 2.3.19, we can introduce another fundamental notion.

**Definition 2.3.20.** Let $F \subset \Pi$ and $G$ a Gröbner basis for $\langle F \rangle$. Then the unique remainder $r$ from Algorithm 2.3.1, starting from some $h \in \Pi$ is called the NORMAL FORM of $h$ with respect to $G$, in short $\nu_G(h)$.

**Remark 2.3.21.** The normal form depends, strictly speaking, not only on $h$ and $G$, but also on the grading. Since term orders always use $\Gamma = \mathbb{N}_0^s$, the chosen term order becomes relevant and indeed the normal form is usually (but not always) strongly dependent on the term order. Therefore, speaking of a Gröbner basis always includes the assumption that the parameters of the grading have been fixed appropriately.

**Corollary 2.3.22.** *If $G$ is a Gröbner basis[22] then*

$$f \in \langle G \rangle \qquad \Leftrightarrow \qquad \nu_G(f) = 0.$$

*Since $\nu_G(f)$ can be computed, the ideal membership problem is decidable.*

### 2.3.4 Division with remainder: a naive only algorithm

Our next goal is to extend the division algorithm 2.3.1 to arbitrary gradings of $\Pi$ and to define the respective generalization of Gröbner bases. This is indeed straightforward since we only have to plugin different notions of degree.

**Definition 2.3.23** ($\Gamma$-basis)**.** A finite[23] set $G \subset \Pi$ is called a $\Gamma$-BASIS for the ideal $\mathscr{I}$ with respect to the grading $\Gamma$ if

$$f \in \mathscr{I} \qquad \Rightarrow \qquad f = \sum_{g \in G} f_g\, g, \quad \delta_\Gamma(f) \geq \delta_\Gamma(f_g g),\ g \in G. \tag{2.3.20}$$

In the case of a term order, we also call it a GRÖBNER BASIS, for the homogeneous grading by total degree an H-BASIS.

**Remark 2.3.24.** Recently, also the name MACAULAY BASIS for an H-basis has become popular in the literature since theses bases were introduced by MACAULAY in [Macaulay, 1916] even long before the invention of Gröbner bases in [Buchberger, 1965], cf. [Buchberger, 1985, Buchberger, 1998] for some historical myths and legends on Gröbner bases. But since the name "H-basis" which comes from a concept of homogenization and dehomogeneization, cf. [Gröbner, 1970, Möller and Sauer, 2000a], was good enough for Macaulay and Gröbner, there seems no real reason or justification for a different terminology.

Unfortunately, Algorithm 2.3.1 cannot be extended directly since it relies fundamentally on properties of monomials:

1. monomials either divide each other or not, and in the latter case they are significantly different. Divisibility is a clear, simple and well–defined property. Classical homogeneous polynomials, almost never divide each other and on the other hand are almost never totally different. As an example consider the (homogeneous) form $x^2 + y^2$ and the form $x^3 + y^3$. No divisibility at all, but still quite a bit in common.

2. For any monomial $x^\alpha$, $0 \neq \alpha \in \mathbb{N}_0^n$ there always exists at least one other monomial $x^\beta$, dividing this one, any nonzero multiindex is contained in several upper sets. Therefore, there is always a chance to eliminate monomials if the divisor set is appropriate. Also this property is lost for homogeneous polynomials where, for example $x^2 + y^2$ is no more divisible by linear forms.

[22] Which always means *"a Gröbner basis for the ideal it generates"*.

[23] Finiteness is not really needed, we add it for convenience since the Basissatz tells us anyway that bases can be chosen finite and in fact we will find out that starting with a finite basis will end with a finite $\Gamma$-basis.

The solution to this problem given in [Sauer, 2001] is to introduce a *gradual* notion of divisibility for homogeneous elements of an arbitrary grading that does not simply distinguish between "divisible" and "not divisible" but considers a concept of "more or less divisible". If, in addition, one aims for numerical stability, for example when the coefficients are only given as FLOATING POINT numbers, then orthogonality is almost self–suggesting.

To be able to handle inner products appropriately, we suppose that the underlying field $\mathbb{K}$ is embedded in $\mathbb{C}$ which essentially means $\mathbb{K} = \mathbb{Q}$, some algebraic extension of $\mathbb{Q}$, $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$. An INNER PRODUCT is then a nondegenerate sesquilinear form

$$\begin{array}{rclcrcl} (f+f',g) & = & (f,g)+(f,g), & & (\lambda f,g) & = & \lambda\,(f,g) \\ (f,g+g') & = & (f,g)+(f,g'), & & (f,\lambda g) & = & \overline{\lambda}\,(f,g) \\ (f,g) & = & \overline{(g,f)} & & (f,f) & \neq & 0. \end{array}$$

To obtain a reasonable Hilbert space we have to require $(f,f) > 0$, which we usually will ensure, but for our purposes it would actually suffice if the inner product were DEFINITE as then

$$W \subset \Pi \qquad \Rightarrow \qquad W \cap W^{\perp} = \{0\}, \qquad W^{\perp} := \big\{ f \in \Pi : (f,W) = 0 \big\}. \tag{2.3.21}$$

Recall also the inner products we already mentioned in (2.1.6) or (2.1.7). Once we have an inner product, hence orthogonality, we can build a relationship between the inner product and the grading to obtain a generalized concept of divisibility.

**Definition 2.3.25.** Let $\Gamma$ be a GRADING[24] and $(\cdot,\cdot) : \Pi \times \Pi \to \mathbb{K}$ and inner product of polynomials.

1. For a finite set $F \subset \Pi$ and $\gamma \in \Gamma$ we denote by

$$V_\gamma(F) := \left\{ \sum_{f \in F} g_f\, \lambda(f) \,:\, g_f \in \Pi_{\gamma - \delta(f)} \right\} \subseteq \Pi_\gamma \tag{2.3.22}$$

   the homogeneous subspace generated by the leading parts of $\lambda(F)$ in $\Pi_\gamma$ and hence also in $\Pi$.

2. In (2.3.22) we use the convention that $\Pi_{\gamma-\eta} = \{0\}$ if $\gamma - \eta$ is undefined in $\Gamma$, that is, if $\eta \notin \gamma + \Gamma$.

3. For $\gamma \in \Gamma$ we denote by

$$W_\gamma(F) := V_\gamma^{\perp}(F) := \Pi_\gamma \ominus V_\gamma(F) = \big\{ g \in \Pi_\gamma \,:\, \big(g, V_\gamma(F)\big) = 0 \big\}$$

   the ORTHOGONAL COMPLEMENT of $V_\gamma(F)$ in $\Pi_\gamma$ and write

$$V(F) = \bigoplus_{\gamma \in \Gamma} V_\gamma(F), \qquad W(F) = \bigoplus_{\gamma \in \Gamma} W_\gamma(F)$$

   belong to the $V_\gamma(F)$, $\gamma \in \Gamma$, or to their orthogonal components.

4. We say that $F \subset \Pi$ DIVIDES $g \in \Pi$ with REMAINDER $r \in \Pi$, if

$$g = \sum_{f \in F} g_f\, f + r, \qquad r \in W(F). \tag{2.3.23}$$

[24] That is, a grading monoid equipped with a well ordering.

**Remark 2.3.26** (Orthogonality & grading)**.**

1. $V_\gamma(F)$ is a $\mathbb{K}$ vector space since any grading must satisfy $\mathbb{K} \subseteq \Pi_0$ and therefore $g \in \Pi_{\gamma-\delta(f)}$ implies that $\mathbb{K} \cdot g \subseteq \Pi_{\gamma-\delta(f)}$.

2. In the case of a term order the concepts from Definition 2.3.25 do not give anything new. Since for $\beta \in \mathbb{N}_0^n$

$$V_\beta(F) = \begin{cases} \Pi_\beta, & \beta \in \bigcup_{f\in F} \left(\delta(f) + \mathbb{N}_0^n\right), \\ \{0\}, & \beta \notin \bigcup_{f\in F} \left(\delta(f) + \mathbb{N}_0^n\right), \end{cases}$$

   see Fig. 2.3.2, we have $r \in W(F)$ if and only if no term from $\lambda(F)$ divides any term $r$, hence, $\operatorname{supp} r \subseteq \mathbb{N}_0^s \setminus U(\delta(F))$ as before.

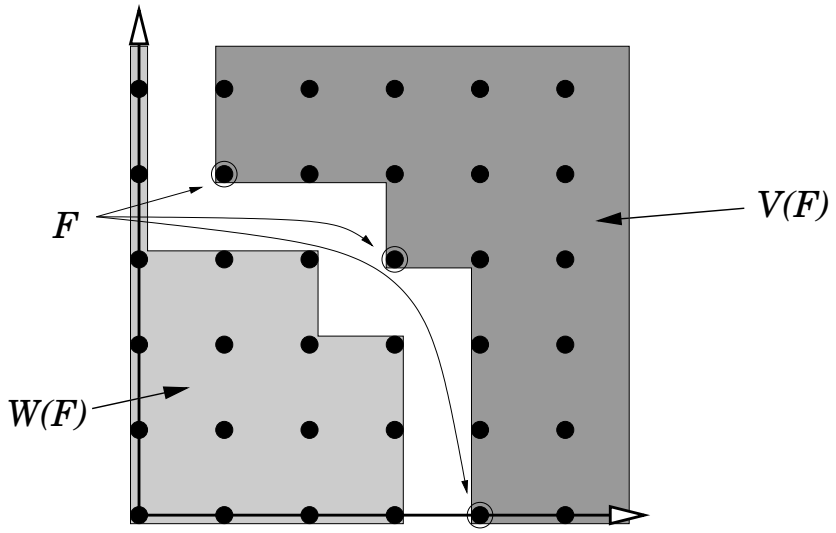



Figure 2.3.2: A finite part of the sets $V(F)$ and $W(F)$ for $F$ consisting of the three monomials $xy^4$, $x^3y^3$ and $x^4$.

3. While for term orders the inner product plays no role, this is not true for general gradings. It turns out that, not as a big surprise, good choices are the inner products from (2.1.6) and (2.1.7).

**Example 2.3.27.** With the H-grading by homogeneous forms we can provide a more interesting example for this concept. Considering, for example, $F = \{x^2 + y^2\}$, the inner product product $(f, f') = f(D)f'(0)$ and $g = x^3 + y^3$, then the first relevant $V_3(F)$ has the form

$$V_3(F) = \operatorname{span}_{\mathbb{K}} \{x^3 + xy^2, x^2y + y^3\},$$

and the basis elements elements are even orthogonal. This implies

$$W_3(F) = \operatorname{span}_{\mathbb{K}} \{x^3 - 3xy^2, 3x^2y - y^3\}$$

and the decomposition is

$$g(x) = \underbrace{\left(\frac{3}{4}x + \frac{3}{4}y\right)\left(x^2 + y^2\right)}_{\in V_3(F)} + \underbrace{\left(\frac{1}{4}x^3 - \frac{3}{4}x^2y - \frac{3}{4}xy^2 + \frac{1}{4}y^3\right)}_{\in W_3(F)}.$$

Now we have all tools available to give a more general version of Algorithm 2.3.1 for arbitrary gradings in Algorithm 2.3.2 and then also to advance the theory.

**Algorithm 2.3.2** DIVISION WITH REMAINDER; $g \in \Pi$, $F \subset \Pi$, grading $\Gamma$

1: $r \leftarrow 0, \qquad g_f \leftarrow 0, f \in F.$
2: **while** $g \neq 0$ **do**
3: $\gamma \leftarrow \delta(g), \qquad h \leftarrow \lambda(g)$
4: *Homogeneous orthogonal projection:* find $h_f \in \Pi_{\gamma-\delta(f)}$, $f \in F$, such that

$$V_\gamma(F) \perp r' := h - \sum_{f\in F} h_f \lambda(f) \tag{2.3.24}$$

5: Set

$$g \leftarrow g - \sum_{f\in F} h_f f - r', \qquad r \leftarrow r + r', \qquad g_f \leftarrow g_f + h_f, \quad f \in F \tag{2.3.25}$$

6: **end while**
7: Result:

$$g = \sum_{f\in F} g_f f + r, \qquad \delta(g) \geq \delta(g_f f), \delta(r), \qquad r \in W(F). \tag{2.3.26}$$

For formalism, we extend the concept of a G-representation, which was attached exclusively to a term order, to more general gradings.

**Definition 2.3.28** ($\Gamma$- and H-representations)**.** A $\Gamma$-REPRESENTATION of $g \in$ with respect to a finite set $F$ is an expression of the form

$$g = \sum_{f\in F} g_f f + r, \qquad \delta_\Gamma(g) \geq \delta_\Gamma(g_f f), \delta_\Gamma(r). \tag{2.3.27}$$

In the case of the HOMOGENEOUS GRADING by total degree, we also call (2.3.27) a H-REPRESENTATION

**Proposition 2.3.29.** *Algorithm 2.3.2 terminates after finitely many steps and determines a* $\Gamma$*-representation* (2.3.26) *with the additional property that* $r \in W(F)$.

**Proof:** The proof closely resembles that of Proposition 2.3.16 and like there we denote by $g_j$ the value of $g$ in the $j$th step of the algorithm, initialized with $g_0 = g$. Since the *homogeneous factors* $h_f$ are chosen such that $\delta\left(h_f f\right) = \delta\left(h_{f'} f'\right)$, $f, f' \in F$, we have that

$$\lambda\left(\sum_{f\in F} h_f f - r_j\right) = \sum_{f\in F} h_f \lambda(f) - \lambda\left(r_j\right) = h_j = \lambda\left(g_j\right),$$

and the degree reduces in each step, that is $\delta\left(g_{j+1}\right) < \delta\left(g_j\right)$.

Correctness, i.e., the validity of (2.3.26) again follows from the invariance

$$g = g_j + \sum_{f\in F} g_f f + r, \qquad j = 0, 1, 2, \ldots,$$

which trivially holds for $j = 0$ by setup and which can be proved inductively by noticing that

$$\begin{aligned} g_{j+1} &= g_j - \sum_{f\in F} h_f f - r_j = g - \sum_{f\in F} g_f f - r - \sum_{f\in F} h_f f - r_j \\ &= g - \sum_{f\in F} \underbrace{\left(g_f + h_f\right)}_{=:\tilde{g}_f} f - \underbrace{r + r_j}_{=:\tilde{r}} \end{aligned}$$

where $\tilde{g}_f$ and $\tilde{r}$ stand for the updated values after performing (2.3.25). □

Of course, we cannot expect some general uniqueness of the Γ-representation or the remainder for more general gradings if this already fails for any term order. However, things are different for a Γ-basis. So we now give the generalized version of Theorem 2.3.19, this time even with a proof.

**Theorem 2.3.30.** *For a* Γ*-basis* $G$ *and any two* Γ*-representations*

$$f = \sum_{g\in G} f_g\, g + r = \sum_{g\in G} f'_g\, g + r', \qquad r, r' \in W(G),$$

*of the same function* $f$ *one has* $r = r'$.

**Proof:** Suppose $r \neq r'$. Since the polynomial

$$0 \neq q := r - r' = \sum_{g\in G} \left(f'_g - f_g\right) g$$

belongs to the ideal $\mathscr{I} = \langle G\rangle$, it has a Γ-representation

$$q = \sum_{g\in G} q_g\, g, \qquad \delta\left(q_g\, g\right) \leq \delta(q) \tag{2.3.28}$$

with respect to the Γ-basis $G$. Considering the leading parts in (2.3.28), we find that

$$0 \neq \lambda(q) = \sum_{g\in G'} \lambda\left(q_g\right) \lambda(g) \in V_{\delta(q)}(G), \qquad G' = \left\{g \in G : \lambda\left(q_g\, g\right) = \lambda(q)\right\} \neq \emptyset.$$

In particular, $\lambda(q) \in V(G)$. On the other hand the remainders have the property that $r, r' \in W(G)$, hence $q = r - r' \in W(G)$ and therefore $\lambda(q) \in W(G)$. This yields

$$\lambda(q) \in W(G) \cap V(G) = \{0\},$$

contradicting $q \neq 0$. □

To close the section we finally give, in generalization of Definition 2.3.20, the concept of a normal form for arbitrary gradings.

**Definition 2.3.31.** Let $G$ be a Γ-Basis of $\mathscr{I}$. Then the NORMAL FORM of $f \in \Pi$ with respect to $G$ or $\mathscr{I}$, written as $\nu_G(f)$ or $\nu_{\mathscr{I}}(f)$, is defined as the remainder of the division in Algorithm 2.3.2.

To show that the notion $\nu_{\mathscr{I}}$ makes sense and really depends on the ideal $\mathscr{I}$ only, we have to show that any two potentially different Γ-bases lead to the same remainder. So let us do that.

**Lemma 2.3.32.** *If* $G, G'$ *are both a* Γ*-basis for the ideal* $\mathscr{I} = \langle G\rangle = \langle G'\rangle$*, then* $\nu_G(f) = \nu_{G'}(f)$, $f \in \Pi$.

**Proof:** For $g' \in G' \subset \langle G \rangle$ let

$$g' = \sum_{g \in G} h_{g',g}\, g, \qquad \delta\left(h_{g',g}\, g\right) \leq \delta\left(g'\right),$$

be the $\Gamma$-representation of $g$ with respect to the $\Gamma$-basis $G$, and let

$$f = \sum_{g' \in G'} f_{g'}\, g' + \nu_{G'}(f), \qquad \delta\left(f_{g'}\, g'\right) \leq \delta(f),$$

be the $\Gamma$-representation of $f$ with respect to $G'$. Then

$$\begin{aligned} f &= \sum_{g' \in G'} f_{g'}\, g' + \nu_{G'}(f) = \sum_{g' \in G'} f_{g'} \left( \sum_{g \in G} h_{g',g}\, g \right) + \nu_{G'}(f) \\ &= \sum_{g \in G} \underbrace{\left( \sum_{g' \in G'} f_{g'}\, h_{g',g} \right)}_{=: f_g} g + \nu_{G'}(f) \end{aligned}$$

is also a $\Gamma$-representation since

$$\begin{aligned} \delta\left(f_g\, g\right) &= \delta(g) + \delta\left( \sum_{g' \in G'} f_{g'}\, h_{g',g} \right) \leq \delta(g) + \max_{g' \in G'} \left(\delta\left(h_{g',g}\right) + \delta\left(f_{g'}\right)\right) \\ &\leq \max_{g' \in G'} \left(\delta(g) + \delta\left(h_{g',g}\right) + \delta\left(f_{g'}\right)\right) \leq \max_{g' \in G'} \left(\delta\left(g'\right) + \delta\left(f_{g'}\right)\right) \leq \delta\left(f\right), \end{aligned}$$

and Theorem 2.3.30 allows us to conclude that $\nu_G(f) = \nu_{G'}(f)$. □

**Remark 2.3.33.** To be precise, the normal form depends on the ideal and on the grading parameters, in particular the inner product used in Algorithm 2.3.2. And indeed, different inner products can lead to different remainders. Since we never vary the inner product "on the fly" and always consider it fixed, this is not a problem, however.

Once we have a unique normal form, we can talk about quotient spaces.

**Definition 2.3.34.** For an ideal $\mathscr{I} \subset \Pi$, we define the associated QUOTIENT SPACE $\Pi/\mathscr{I}$ as $\nu_{\mathscr{I}}(\Pi)$ with the operations

$$f + f' := \nu_{\mathscr{I}}(f + f'), \quad f \cdot f' := \nu_{\mathscr{I}}(f \cdot f'), \qquad f, f' \in \Pi/\mathscr{I}. \tag{2.3.29}$$

**Remark 2.3.35.** Defining the addition in (2.3.29) is notational overkill since $\Pi/\mathscr{I}$ is always a $\mathbb{K}$ vector space due to the linearity of the division process. But it is also not wrong.

**Exercise 2.3.1** Show that $\nu_{\mathscr{I}}(f + f') = \nu_{\mathscr{I}}(f) + \nu_{\mathscr{I}}(f')$, $f, f' \in \Pi$. ♢

## 2.4 Computing good bases

In the preceding chapter we have seen that $\Gamma$-bases would be an extremely useful tool because they would allow us to extend the concept of division with remainder to multivariate polynomials, thus making the non-euclidean ring some sort of unique. This leads to two fundamental questions:

1. For which ideals $\mathscr{I} \subseteq \Pi$ and for which gradings does there exist a $\Gamma$-basis? Do they exist at all?
2. If they exist, can we efficiently construct such a basis, for example from another given basis, i.e., given $F$ can we find a $\Gamma$-basis $G$ such that $\langle F\rangle = \langle G\rangle$.

Fortunately the answer to both questions is "yes", so let us also try to convince ourselves by proving that fact.

### 2.4.1 Good bases and division

The fundamental tool used in the construction of $\Gamma$-bases is the following characterization by means of the division algorithm 2.3.2.

**Proposition 2.4.1.** *A finite set $G \subset \Pi$ is a $\Gamma$-basis if and only for any $q = (q_g : g \in G) \in \Pi^G$ such that*

$$\delta(q \cdot G) := \delta\left(\sum_{g\in G} q_g\, g\right) < \max_{g\in G} \delta(q_g\, g) \tag{2.4.1}$$

*the division algorithm 2.3.2 gives the remainder*

$$r = \nu_G(q \cdot G) = 0. \tag{2.4.2}$$

**Proof:** If $G$ is a $\Gamma$-basis then the polynomial $f := q \cdot G \in \langle G\rangle$ has, by Definition 2.3.23 a $\Gamma$-representation with respect to $G$ and remainder 0, which, by Theorem 2.3.30 implies that $\nu_G(q \cdot G) = 0$.

For the direction "$\Leftarrow$" we use an approach that was given in [Möller, 1988] for the construction of a Gröbner basis. Denoting by $r_G(f)$, $f \in \Pi$, the remainder obtained in Algorithm 2.3.2, we thus assume that $r_G(q \cdot G) = 0$ whenever $\delta(q \cdot G) < \max \delta(q_g\, g)$, $q \in \Pi^G$. Any $f \in \langle G\rangle$ has, by definition, a representation

$$f = \sum_{g\in G} f_g\, g, \qquad f_g \in \Pi, \quad g \in G. \tag{2.4.3}$$

which need not be a $\Gamma$-representation and our goal will be to transform it into

$$f = \sum_{g\in G} f'_g\, g, \qquad \delta\left(f'_g\, g\right) \le \delta(f), \quad g \in G,$$

as that would prove that $G$ is a $\Gamma$-basis. Supposing that (2.4.3) is *no* $\Gamma$-representation, we set

$$\gamma := \max_{g\in G} \delta(f_g\, g) > \delta(f) \qquad \text{and} \qquad J := \{g \in G : \delta(f_g\, g) = \gamma\},$$

so that the specific choice of $q$ as

$$q = (q_g : g \in G) \qquad \text{mit} \qquad q_g = \begin{cases} \lambda(f_g), & g \in J, \\ 0 & g \notin J, \end{cases} \quad g \in G, \tag{2.4.4}$$

has the property

$$\delta(q \cdot G) < \gamma = \max_{g\in G} \delta(q_g\, g).$$

By the division algorithm and our assumptions on $G$ we thus get a $\Gamma$-representation

$$q \cdot G = \sum_{g\in G} q'_g\, g + \underbrace{r_G(q \cdot G)}_{=0} = \sum_{g\in G} q'_g\, g, \qquad \delta\left(q'_g\, g\right) < \gamma.$$

Thus,

$$\begin{aligned} f &= \sum_{g\in G} f_g\, g = \sum_{g\in G\setminus J} f_g\, g + \sum_{g\in J} \left(f_g - \lambda\left(f_g\right)\right) g + \underbrace{\sum_{g\in J} \lambda\left(f_g\right) g}_{=q\cdot g = q'\cdot g} \\ &= \sum_{g\in G\setminus J} f_g\, g + \sum_{g\in J} \left(f_g - \lambda\left(f_g\right)\right) g + \sum_{g\in G} q'_g\, g \qquad (2.4.5) \\ &=: \sum_{g\in G} f^1_g\, g \end{aligned}$$

and since all representations in (2.4.5) have degree $< \gamma$, we can conclude that

$$\delta(f) \le \delta\left(f^1_g\, g\right) < \gamma, \qquad g \in G.$$

We proceed in the same way, set $\gamma_1 := \max_g \delta\left(f^1_g\, g\right)$ and if still $\gamma_1 > \delta(f)$, we use the same argument to obtain $\gamma_2 < \gamma_1$ with associated coefficients $f^2_g$, $g \in G$, and so on. Since a grading is always based on a well ordering by definition, this process must terminate after finitely many, say $N$, steps, where $\gamma_N = \delta(f)$ as otherwise we could reduce the degree even further. But then $f'_g := f^N_g$, $g \in G$, yields a $\Gamma$-representation for $f$ which proves that $G$ is a $\Gamma$-basis. □

**Remark 2.4.2.** It is worthwhile to note that Proposition 2.4.1 is only a property of degrees and grading. That we used orthogonal projections when needed, makes the algorithm work but does not affect this result in any way.

The proof of Proposition 2.4.1 shows us that we can even get a stronger statement: in the crucial step (2.4.4) we even chose the vector $q$ in such a way that all its entries were homogeneous polynomials, i.e., we not only had $q \in \Pi^G$ but even $q \in \left(\Pi^0\right)^G$. The requirement $\delta\left(q\cdot G\right) < \gamma$ then means that

$$q \cdot \lambda(G) = \sum_{g\in J} q_g\, \lambda(g) = 0. \qquad (2.4.6)$$

Such tuples have a cute name and a story of their own which will be told in the next subsection.

### 2.4.2 Syzygies

Syzygies play a fundamental in the study of polynomial ideals and related problems. So let us define what this is.

**Definition 2.4.3.** Let $F \subset \Pi$ be finite.

1. $s \in \Pi^F$ is called a SYZYGY $s \cdot F = 0$.
2. A module over a ring $R$ is a set that is closed under addition and multiplication with elements of $R$.
3. The SYZYGY MODULE of a set $F \subset \Pi$ will be denoted by

$$S(F) = \left\{ s \in \Pi^F \,:\, 0 = s\cdot F := \sum_{f\in F} s_f\, f \right\}.$$

**Remark 2.4.4.** The name *syzygy* is composed from the Greek words "$\sigma\upsilon\varsigma$" = "together" and "$\zeta\upsilon\gamma o\upsilon$" = "yoke" and originally means something bound together like two oxen pulling a cart. The word was already used in Greek astrononomy, its Latin translation is "coniunctio" which lives on in "conjunction". The word "syzygy" has one more meaning that is pointed out in [Eisenbud, 1994].

**Remark 2.4.5** (Simple module properties)**.**

1. $S(F)$ is a module over $\Pi$, as for $s, s' \in S(F)$ and $g, g' \in \Pi$ we have that

$$0 = g\,(s \cdot F) + g'\left(s' \cdot F\right) = (g\,s + g\,s') \cdot F,$$

   hence $g\,s + g'\,s' \in S(F)$.

2. Any ideal is a module, tuples of polynomials form modules.

3. Intuitively modules can be interpreted as some sort of "vector spaces over rings". This is why we use the dot notation for syzygies that reminds of an inner product.

4. Vector spaces are modules over fields.

5. The module $S(F)$ is FINITELY GENERATED, see [Gröbner, 1970]. This means that there exists a *finite* set $S \subset S(F)$ such that

$$S(F) = \left\{ \sum_{s \in S} q_s\, s \,:\, q_s \in \Pi \right\}. \tag{2.4.7}$$

**Definition 2.4.6.** A set $S$ such that (2.4.7) holds is called a BASIS for the syzygy module $S(F)$.

The next lemma characterizes $\Gamma$-bases in terms of syzygies. It looks innocent, but is the key tool for their construction by division with remainder. In Gröber basis terms it says that *"every syzygy of leading terms must reduce to zero"*.

**Lemma 2.4.7.** *Let $G \subset \Pi$ be finite and $S \in S(\lambda(G))$ be a basis for the syzygy module of leading terms of $G$. Then $G$ is a $\Gamma$-basis if and only if $r_G(s \cdot G) = 0$, $s \in S$.*

**Proof:** According to Proposition 2.4.1, $G$ is a $\Gamma$-basis if and only if $r_G(s \cdot G) = 0$ for all $s \in S(\lambda(G))$. So all we have to show is that is suffices to consider a basis of the syzygy module.

If already $r_G(s \cdot G) \neq 0$ for some $s \in S$, then $r_G(s \cdot G) = 0$, $s \in S(\lambda(F))$, is trivially impossible. If conversely

$$S(\lambda(F)) \ni t = \sum_{s \in S} q_s\, s, \qquad q_s \in \Pi,$$

is an arbitrary syzygy of leading parts, then the assumption that any element of $S$ reduces to zero leads to

$$t \cdot G = \sum_{g \in G} t_g\, g = \sum_{g \in G} \sum_{s \in S} q_s s_g\, g = \sum_{s \in S} q_s \sum_{g \in G} s_g\, g = \sum_{s \in S} q_s\,(s \cdot G)$$

and

$$r_G(t \cdot G) = \sum_{s \in S} r_G\left(q_s\,(s \cdot g)\right) = 0 \sum_{s \in S} q_s\, r_G(s \cdot g) = 0,$$

since the division algorithm factorized of multiples. □

### 2.4.3 Buchberger's algorithm

Lemma 2.4.7 already hints how to construct a $\Gamma$-basis by successively enlarging or completing a given basis of the ideal until all syzygies of leading terms reduce to zero:

1. Determine a finite basis $S$ of $S(\lambda(F))$.
2. If $r_F(s \cdot F) = 0$ for all $s \in S$, then, by Lemma 2.4.7, the set $F$ is a $\Gamma$-basis and we are done.
3. If not, there must be an $s^* \in S$ such that
$$0 \neq f^* := r_F\left(s^* \cdot F\right) = \underbrace{s^* \cdot F}_{\in\langle F\rangle} - \underbrace{\sum_{f\in F} g_f f}_{\in\langle F\rangle} \quad \in \langle F\rangle.$$
4. By this observation the enlarged basis still maintains $\left\langle F \cup \left\{f^*\right\}\right\rangle = \langle F\rangle$, but also has the property that
$$r_{F\cup\{f^*\}}\left(s^* \cdot F\right) = 0,$$
and therefore now reduces the bad syzygy to zero. In other words: enlarging the basis makes is more $\Gamma$ish.
5. Repeat the process with $F \cup \left\{f^*\right\}$ instead of $F$.

In principle this is already the $\Gamma$-version of Buchberger's algorithm which was developed by in 1965 by Bruno Buchberger, [Buchberger, 1965] in a PhD thesis supervised by Gröbner.

**Algorithm 2.4.1** BUCHBERGER'S ALGORITHM: $F \subset \Pi$ finite basis of $\mathscr{I}$.

**repeat**
  Determine finite basis $S$ of $S(\lambda(F))$.
  $G \leftarrow \{r_F(s \cdot F) : s \in S\} \setminus \{0\}$.
  $F \leftarrow F \cup G$.
**until** $G = \emptyset$.
Result: $\Gamma$-basis $F$ for $\mathscr{I}$.

And yes, it works.

**Theorem 2.4.8.** *Algorithm 2.4.1 terminates after finitely many steps and yields a $\Gamma$-basis for the ideal.*

**Remark 2.4.9.** Algorithm 2.4.1 relies on the construction of a finite basis for the syzygies of leading terms. We will see very soon that this is easy and can be done explicitly whenever one considers only a term order. This is the reason why Gröbner bases are by far the most popular computational ideal bases, in particular as different term orders provide quite some flexibility. Moreover, once a Gröbner basis has been computed for an ideal, it is possible to compute a basis of the module of syzygies, see [Buchberger, 1985]. Conceptionally, however, this is not satisfactory.

**Lemma 2.4.10.** *Let $F \subset \Pi$ be finitely and $\Gamma$ a term order. The the syzygy module $S(\lambda(F))$ of the leading terms of $F$ is generated by the* S-POLYNOMIALS *$s\left(f, f'\right)$, $f, f' \in F$, $f \neq f'$, whose nonzero components are defined as*
$$s\left(f,f'\right)_f = \frac{\lambda\left(f'\right)}{x^\alpha}, \qquad s\left(f,f'\right)_{f'} = -\frac{\lambda(f)}{x^\alpha}, \qquad \alpha = \min\left\{\delta(f), \delta\left(f'\right)\right\}, \tag{2.4.8}$$
*where the minimum in (2.4.8) has to be understood in a componentwise sense.*

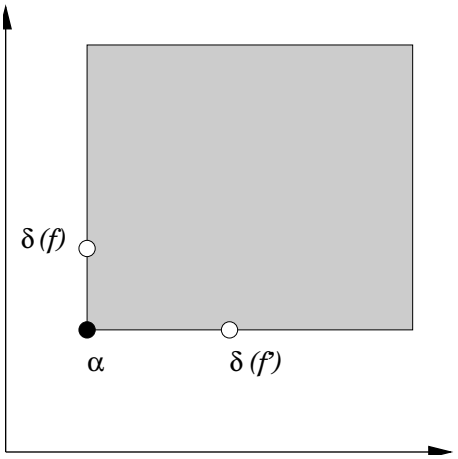



Figure 2.4.3: The geometric interpretation of $\alpha$ in (2.4.8) as maximal multiindex that generates a cone or upper set $\alpha+\mathbb{N}_0^s$ which contains $\delta(f)$ as well as $\delta\left(f'\right)$. In this sense, $x^\alpha$ can be understood as $\gcd\left(\lambda\left(f\right),\lambda\left(f'\right)\right)$.

S-polynomials are the simplest form of syzygies imaginable as they are syzygies among two polynomials that can easily be computed by hand. That they nevertheless generate all other syzygies is a lucky accident but very special for the case of a term order.

**Proof of Lemma 2.4.10:** Since any syzygy can be decomposed into its homogeneous components[25], i.e., into terms, we can assume that $s \in S(F)$ is a HOMOGENEOUS SYZYGY, that is, there exists $\beta \in \mathbb{N}_0^n$ such that

$$s_f\,\lambda\left(f\right) \in \Pi_\beta, \qquad \text{also} \qquad s_f\,\lambda\left(f\right) = c_j\,x^\beta, \qquad f \in F.$$

If $s \in S(F)$ is a nontrivial syzygy where two nonzero terms sum to zero, then there exist at least *two* polynomials $f, f' \in F$ as indices such that $s_f, s_{f'} \neq 0$. In particular,

$$\beta \in \left(\delta(f)+\mathbb{N}_0^n\right) \cap \left(\delta\left(f'\right)+\mathbb{N}_0^n\right) \subset \alpha+\mathbb{N}_0^n.$$

There exists a term $g' \in \Pi_{\beta-\delta(f')}$ with

$$s_f\,\lambda(f) = g'\,\lambda\left(f'\right) = d\,x^\beta, \qquad d \in \mathbb{K},$$

which allows us to conclude that

$$s\cdot\lambda(F) = \sum_{h\in F\setminus\{f\}} s_h\lambda(h) + \underbrace{s_f\lambda(f) - g'\,\lambda\left(f'\right)}_{=g\,s_{f,f'}} + g'\,\lambda\left(f'\right) =: s'\cdot\lambda(F) + \left(g\,s_{f,f'}\right)\cdot\lambda(F),$$

and consequently $s' = s - g\,s_{f,f'} \in S(F)$ has one less nonzero entry than $s$. Iterating this procedure, we eventually obtain the decomposition into S-polynomials. □

To prove termination of Buchberger's algorithm we need one more characterization of Γ-bases, based on the concept of a HOMOGENEOUS IDEAL which is an ideal $\mathscr{H}$ such that $f \in \mathscr{H}$ implies that all homogeneous components of $f$ belong to the ideal again:

$$\mathscr{H} \ni f = \sum_{\gamma\in\Gamma} f_\gamma \qquad \Rightarrow \qquad f_\gamma \in \mathscr{H}, \quad \gamma \in \Gamma. \tag{2.4.9}$$

[25] A polynomial is the zero polynomial if and only if all its homogeneous components are zero.

This concept has been defined in [Gröbner, 1970] for the classical homogeneous grading by total degree, in the context of a term order it is nothing but a monomial ideal. Indeed, if $\mathscr{H}$ has a basis $F \subset \Pi^0$ of $\Gamma$-FORMS[26], the $\mathscr{H}$ obviously is a homogeneous ideal. If, conversely, $\mathscr{H}$ is a homogeneous ideal and $F$ a basis for $\mathscr{H} = \langle F \rangle$, then the set

$$F^0 := \{ f_\gamma \,:\, f \in F, \gamma \in \Gamma \} \subset \Pi^0$$

of homogeneous parts appearing in $F$ belongs to $\mathscr{H}$ as well, generates[27] $F \subseteq \langle F^0 \rangle$ and therefore

$$\mathscr{H} = \langle F \rangle \subseteq \langle F^0 \rangle.$$

We can summarize this as follows.

**Lemma 2.4.11.** *An ideal $\mathscr{H} \subset \Pi$ is a homogeneous ideal if and only if it has a basis consisting of homogeneous polynomials.*

**Remark 2.4.12.** A monomial ideal is exactly a homogeneous ideal if the underlying grading is by term order. Otherwise it is an ideal generated by ($\Gamma$-)forms which explains the old German name FORMENIDEAL used for such ideals in [Gröbner, 1970].

**Remark 2.4.13.** There is a different approach to homogeneous ideals as ideals where multiplication is allowed only for homogeneous polynomials and addition only for homogeneous polynomials of the same degree.

The last observation is once more very easy to prove but fundamental for the termination of Buchberger's algorithm.

**Lemma 2.4.14.** *For any ideal $\mathscr{I} = \langle G \rangle \subset \Pi$ one has:*

1. *the ideals $\langle \lambda(G) \rangle$ and $\langle \lambda(\mathscr{I}) \rangle$ are each a homogeneous ideal.*
2. *$G$ is a $\Gamma$-basis for $\mathscr{I}$ if and only if $\langle \lambda(\mathscr{I}) \rangle = \langle \lambda(G) \rangle$.*

**Proof:** 1): both ideal are generated by homogeneous polynomials and thus are homogeneous ideals by Lemma 2.4.11. For 2) we assume that $G$ is a $\Gamma$-basis, so that $f \in \mathscr{I}$ has a $\Gamma$-representation

$$f = \sum_{g \in G} f_g \, g,$$

which implies that

$$\lambda(f) = \sum_{\{g \,:\, \delta(f_g\, g) = \delta(f)\}} \lambda\left(f_g\right) \lambda(g) \quad \in \langle \lambda(G) \rangle.$$

If, conversely, $f \in \mathscr{I}$ and $\lambda(f) = h \cdot \lambda(G)$, then $f - h \cdot G$ also belongs to $\mathscr{I}$ but has a strictly smaller degree. Proceeding with the leading part of this polynomial, we iteratively get all homogeneous components of a $\Gamma$-representation of $f$. □

**Proof of Theorem 2.4.8:** Proposition 2.4.1 states that termination of Algorithm 2.4.1, which means that all syzygies of leading parts reduce to zero, indeed classifies $F$ as a $\Gamma$-basis.

It remains to prove termination. To that end, we note that $0 \neq g := r_F(s \cdot F)$ means that[28] $\lambda(g) \in W_{\delta(g)}(F)$, hence $g \in \langle F \rangle$, but $\lambda(g) \notin \langle \lambda(F) \rangle$. Replacing $F$ by $F' = F \cup \{g\}$ it follows that

[26] As generalized homogeneous parts.
[27] The inclusion even holds true in terms of *vector spaces*.
[28] Now we again use our particular form of the reduction algorithm 2.3.2.

$\langle F' \rangle = \langle F \rangle$ but $\langle \lambda(F) \rangle$ is a *proper* subset of $\langle \lambda(F') \rangle$. Since polynomials are a Noetherian ring, such a strictly ascending chain of ideal has to be finite, hence after finitely many steps there cannot be any more $s \in S(F)$ such that $r_F(s \cdot F) \neq 0$, as otherwise we could increase the homogeneous ideal even further. □

**Remark 2.4.15.** Of course, the Buchberger1.0 that we defined in Algorithm 2.4.1 is not really efficient and far from being optimal. How to handle syzygies, which to ignore and which to really compute and how to this in the best possible way, has been an issue ever since the introduction of Gröbner bases in 1965 and there exist a vast literature on this.

### 2.4.4 The Basissatz

The efforts of the preceding chapter, in particular the construction of Gröbner bases allows us to give a proof of Hilbert's Basissatz that we quoted in Theorem 2.1.29. The method is a follows.

**Proof of Theorem 2.1.29:** Let $\mathscr{I}$ be an arbitrary ideal and let $\Gamma = \mathbb{N}_0^s$ stand for the grading by an arbitrary term order. Then the ideal

$$\lambda(\mathscr{I}) := \langle \lambda(f) : f \in \mathscr{I} \rangle$$

generated by all leading terms in $\mathscr{I}$ is a monomial ideal. By Corollary 2.3.10 this ideal is finitely generated, hence there exists a *finite* set $H \subset \Pi$ of monomials such that $\lambda(\mathscr{I}) = \langle G_0 \rangle$. Since any $h \in G_0$ is of the form $h = \lambda(f)$ for some $f \in \mathscr{I}$, we can write $G_0 = \lambda(G)$ for some finite set $G \subset \mathscr{I}$, hence $\langle G \rangle \subseteq \mathscr{I}$ and since

$$\langle \lambda(G) \rangle = \langle G_0 \rangle = \lambda(\mathscr{I}),$$

Lemma 2.4.14 yields that $G$ is even a Gröbner basis for $\mathscr{I}$. Hence, the ideal is finitely generated. □

Since any ideal has a finite basis[29] and since we have Buchberger's algorithms and its generalizations to transform this basis into a $\Gamma$ basis, we can state the following fundamental result.

**Theorem 2.4.16** ($\Gamma$ bases)**.** *For any grading $\Gamma$, any ideal has a finite $\Gamma$-basis.*

### 2.4.5 The homogeneous way

For the homogeneous grading the computation of a basis of the module of syzygies is nontrivial. There is a method in [Buchberger, 1985] that computes a basis for the syzygy module for *any* finite set of polynomials, but to use Gröbner basis especially in a situation where one wants to avoid term orders is contraintuitive. For the homogeneous grading there is an illustrative way to compute syzygies by means of linear algebra which also leads to an efficient way of doing reduction.

**Definition 2.4.17.** Let $f \in \Pi$. By $\boldsymbol{f}_k := \left(f_\alpha : |\alpha| = k\right) \in \mathbb{K}^{r_k}$, $r_k := r_k^s = \binom{k+s-1}{s-1}$ we denote the COEFFICIENT BLOCK of the homogeneous component of degree $k$.

[29] And we have no way to handle ideals other than working on bases.

Any polynomial $f$ of degree $n$ is then represented by the coefficient blocks $\boldsymbol{f}_k$, $k = 0,\dots,n$, and with the MONOMIAL BLOCKS $\boldsymbol{x}^k = (x^\alpha : |\alpha| = k)$, we get the convenient notation

$$f(x) = \sum_{k=0}^{n} \boldsymbol{f}_k^T \boldsymbol{x}^k, \qquad \lambda(f)(x) = \boldsymbol{f}_{\deg f}^T \boldsymbol{x}^{\deg f}, \tag{2.4.10}$$

which has a quite univariate flavor. To that end, we identify $f$ with the block vector

$$f \simeq \boldsymbol{f} = \begin{pmatrix} \boldsymbol{f}_0 \\ \vdots \\ \boldsymbol{f}_{\deg f} \end{pmatrix}. \tag{2.4.11}$$

Next, we represent multiplication by monomials.

**Definition 2.4.18.** The HOMOGENEOUS LIFTING MATRIX $\boldsymbol{L}_{n,k} \in \mathbb{K}^{r_n \times r_k}$, $k \le n$, is defined as

$$\boldsymbol{L}_{n,k} = \left(\boldsymbol{L}_{\alpha,k} : |\alpha| = n-k\right), \qquad \boldsymbol{L}_{\alpha,k} := \sum_{|\beta|=k} \boldsymbol{e}_{\alpha+\beta} \boldsymbol{e}_\beta^T, \quad \alpha \in \mathbb{N}_0^s. \tag{2.4.12}$$

The function of the lifting matrix is easily seen when considering, for a homogeneous polynomial $f \in \Pi_k^0$,

$$\left(\boldsymbol{L}_{\alpha,k} \boldsymbol{f}_k\right)^T \boldsymbol{x}^n = \left( \sum_{|\beta|=k} \boldsymbol{e}_{\alpha+\beta} \boldsymbol{e}_\beta^T \boldsymbol{f}_k \right)^T \boldsymbol{x}^n = \sum_{|\beta|=k} f_\beta \boldsymbol{e}_{\alpha+\beta}^T \boldsymbol{x}^n = \sum_{|\beta|=k} f_\beta x^{\beta+\alpha} = x^\alpha \left( \boldsymbol{f}_k^T \boldsymbol{x}^k \right); \tag{2.4.13}$$

is is simply the multiplication of the homogeneous polynomial by $(\cdot)^\alpha$. In the same way, $\boldsymbol{L}_{n,k} \boldsymbol{f}_k$ generates a matrix whose columns are the coefficient vectors of all multiplications of the homogeneous polynomial $\boldsymbol{f}_k^T \boldsymbol{x}^k$ with all monomials of degree $n-k$.

**Remark 2.4.19.** The computation of $\boldsymbol{V}_n(F)$ is numerically cheap and without roundoff errors since it only consists of redistributing the coefficient vectors.

**Definition 2.4.20.** Given a finite set $F \subset \Pi$ of polynomials, we define the matrix

$$\boldsymbol{V}_n(F) = \left( \boldsymbol{L}_{n,\deg f} \boldsymbol{f}_{\deg f} : f \in F \right) \in \mathbb{K}^{r_n \times N}, \qquad n \in \mathbb{N}_0, \quad N = \sum_{f \in F} r_{n-\deg f}, \tag{2.4.14}$$

and call it the GENERATING MATRIX for $V_n(F) \subset \Pi_n^0$. We use the convention that $\boldsymbol{L}_{n,k}$ is an empty matrix if $n < k$ and $r_k = 0$ for $k < 0$.

**Remark 2.4.21.** Note that $\boldsymbol{V}_n(F)$ is a somewhat complicated object. It is first indexed by $f$, yielding the matrices

$$\left(\boldsymbol{V}_n(F)\right)_f = L_{n,\deg f} \boldsymbol{f}_{\deg f}$$

which are in turn indexed by $\alpha$ to yield the well defined column vector

$$\left( \left(\boldsymbol{V}_n(F)\right)_f \right)_\alpha = \left( \boldsymbol{L}_{n,\deg f} \boldsymbol{f}_{\deg f} \right)_\alpha = \boldsymbol{L}_{\alpha,\deg f} \boldsymbol{f}_{\deg f}.$$

This means that any such matrix can be multiplied from the right hand side with block vectors $\boldsymbol{h}$ indexed as $h_{f,\alpha}$ where

$$\boldsymbol{V}_n(F) \boldsymbol{h} = \sum_{f \in F} \sum_{|\alpha| = n - \deg f} h_{f,\alpha} \boldsymbol{L}_{\alpha,\deg f} \boldsymbol{f}_{\deg f} \in \mathbb{C}^{r_n}.$$

The whole intention of this construction is that the range of $\boldsymbol{V}_n(F)$ consists of the coefficient vectors of the polynomials that span $V_n(F)$.

**Lemma 2.4.22.** *For $n \in \mathbb{N}_0$ and $g \in \Pi_n^0$ we have that*

$$g \in V_n(F) \qquad \Leftrightarrow \qquad \boldsymbol{g}_n \in \boldsymbol{V}_n(F)\,\mathbb{K}^n. \tag{2.4.15}$$

**Proof:** $g \in V_n(F)$ if and only if there exist homogeneous polynomials $h_f = \sum h_{f,\alpha}(\cdot)^\alpha$, $f \in F$, such that

$$g = \sum_{f\in F} h_f\, f = \sum_{f\in F} \sum_{|\alpha|=n-\deg f} h_{f,\alpha}(\cdot)^\alpha f$$

which is, by (2.4.13), equivalent to

$$\begin{aligned} \boldsymbol{g}_n &= \sum_{f\in F} h_{f,\alpha} \sum_{|\alpha|=n-\deg f} \boldsymbol{L}_{\alpha,\deg f}\boldsymbol{f}_{\deg f} = \sum_{f\in F} h_{f,\alpha} \sum_{|\alpha|=n-\deg f} \left(\boldsymbol{L}_{n,\deg f}\boldsymbol{f}_{\deg f}\right)_\alpha \\ &= \sum_{f\in F} \boldsymbol{L}_{n,\deg f}\boldsymbol{f}_{\deg f}\,\boldsymbol{h}_f = \boldsymbol{V}_n(F)\,\boldsymbol{h}, \qquad \boldsymbol{h} = \left(\boldsymbol{h}_f : f \in F\right), \end{aligned}$$

in the sense that $\boldsymbol{h}$ is a concatenation of column vectors. □

Now we can use a standard concept of numerical linear algebra to efficiently perform the reduction step. We embed everything into the most general case $\mathbb{K} = \mathbb{C}$.

**Theorem 2.4.23** (SVD)**.** *For any matrix $A \in \mathbb{C}^{m\times n}$ there exist unitary matrices $U \in \mathbb{C}^{m\times m}$ and $V \in \mathbb{C}^{n\times n}$ and a diagonal matrix $\Sigma \in \mathbb{R}^{m\times n}$ with nonnegative diagonal elements such that*

$$A = U\Sigma V^H. \tag{2.4.16}$$

**Definition 2.4.24** (SVD)**.** The decomposition in (2.4.16) is called SINGULAR VALUE DECOMPOSITION of $A$ or SVD, for short. Each column of $U$ is called a left SINGULAR VECTORS, the columns of $V$ are the right singular vectors. Any diagonal element $\sigma_j$ is called SINGULAR VALUE, the number of nonzero singular values coincides with the rank $r$ of the matrix. Often one uses the THIN SVD

$$A = \sum_{k=1}^{r} \sigma_k u_k v_k^T = \left[U_{(1:r)}\, U_{(r+1:m)}\right] \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} \left[V_{(1:r)}\, V_{(r+1:n)}\right]^H, \qquad \Sigma_r := \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{pmatrix}, \tag{2.4.17}$$

where $\sigma_1 \geq \cdots \geq \sigma_r > 0$, hence $\Sigma_r$ is nonsingular.

**Remark 2.4.25** (SVD)**.**

1. The SVD can be computed efficiently and is one of the most important tools in numerical linear algebra. For some computational details see [Golub and van Loan, 1996].

2. One important application for the SVD is the computation of the KERNEL of $A$, $\ker A := \{x : Ax = 0\}$. In fact, it follows directly from (2.4.17) that the columns of $V_{(r+1:n)}$ are an orthonormal basis for the kernel. The SVD is a RANK REVEALING FACTORIZATION.

3. In the same fashion we can also determine the range of $A$: the columns of $U_{(1:r)}$ form an orthonormal basis for the range and the columns of $U_{(r+1:m)}$ an orthonormal basis for the orthogonal complement of the range in $\mathbb{C}^m$.

4. *Numerically*, the rank of a matrix is determined by computing its SVD and thresholding the singular values; depending on how aggressive this thresholding is performed, the rank may be over- or underestimated.

The observation 3) already shows us how to compute the coefficient vectors for $V_n(F)$ and $W_n(f)$ – an immediate consequence of Remark 2.4.25 and Lemma 2.4.22.

**Lemma 2.4.26.** *Let*

$$\boldsymbol{V}_n(F) = [\boldsymbol{V}_n, \boldsymbol{W}_n]\, \Sigma\, [\boldsymbol{R}_n, \boldsymbol{S}_n]^H \tag{2.4.18}$$

*be a thin SVD of $\boldsymbol{V}_n(F)$, then the columns of $\boldsymbol{V}_n$ are the coefficient vectors of an orthonormal basis of $V_n(F)$ and the columns of $\boldsymbol{W}_n$ are an orthonormal basis of $W_n(F)$.*

This allows us to give an efficient reduction algorithm for the single reduction step in the division with remainder algorithm. To that end, we note that to approximate $\boldsymbol{g}_n$ by $\boldsymbol{V}_n(F)$ we have to find $\boldsymbol{h}$ such that

$$0 = \boldsymbol{V}_n^H \left( \boldsymbol{g}_n - \boldsymbol{V}_n(F)\boldsymbol{h} \right) \tag{2.4.19}$$

that is,

$$\begin{aligned} \boldsymbol{V}_n^H \boldsymbol{g}_n &= \boldsymbol{V}_n^H \boldsymbol{V}_n(F)\boldsymbol{h} = \boldsymbol{V}_n^H\, [\boldsymbol{V}_n, \boldsymbol{W}_n]^H \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} [\boldsymbol{R}_n, \boldsymbol{S}_n]^H \boldsymbol{h} = [\boldsymbol{I}, \boldsymbol{0}] \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} [\boldsymbol{R}_n, \boldsymbol{S}_n]^H \boldsymbol{h} \\ &= [\Sigma_r, \boldsymbol{0}] \begin{pmatrix} \boldsymbol{R}_n^H \\ \boldsymbol{S}_n^H \end{pmatrix} \boldsymbol{h} = \Sigma_r \boldsymbol{R}_n^H \boldsymbol{h}, \end{aligned}$$

hence, choosing $\boldsymbol{h}$ as one solution of the underdetermined system

$$\boldsymbol{R}_n^H \boldsymbol{h} = \Sigma_r^{-1} \boldsymbol{V}_n^H \boldsymbol{g}_n,$$

for example

$$\boldsymbol{h} = \boldsymbol{R}_n \Sigma_r^{-1} \boldsymbol{V}_n^H \boldsymbol{g}_n \tag{2.4.20}$$

does the job since $\boldsymbol{R}_n^H \boldsymbol{R}_n = \boldsymbol{I}_r$ is an $r \times r$ identity matrix. The vector $\boldsymbol{h}$ can be decomposed into

$$\boldsymbol{h} = \left[ \boldsymbol{h}_f : f \in F \right], \qquad \boldsymbol{h}_f \in \mathbb{C}^{r_n - r_{\deg f}},$$

where

$$h_f(x) = \boldsymbol{h}_f^T \boldsymbol{x}^{n-\deg f} = \sum_{|\alpha| = n - \deg f} h_{f,\alpha} x^\alpha$$

is the corresponding polynomial. The coefficients of $h_f f$ are then obtained, again by means of (2.4.13) as

$$\begin{aligned} (h_f f)(x) &= \sum_{|\alpha| = n-\deg f} h_{f,\alpha} x^\alpha \sum_{k=0}^{\deg f} \boldsymbol{f}_k^T \boldsymbol{x}^k = \sum_{k=0}^{\deg f} \sum_{|\alpha| = n-\deg f} x^\alpha \boldsymbol{f}_k^T \boldsymbol{x}^k\, h_{f,\alpha} \\ &= \sum_{k=0}^{\deg f} \sum_{|\alpha| = n-\deg f} \left( L_{\alpha,k} \boldsymbol{f}_k \right)^T \boldsymbol{x}^{k+|\alpha|} h_{f,\alpha} = \sum_{k=0}^{\deg f} \left( \sum_{|\alpha| = n-\deg f} h_{f,\alpha} L_{\alpha,k} \boldsymbol{f}_k \right)^T \boldsymbol{x}^{n-\deg f+k} \\ &= \sum_{k=0}^{\deg f} \left( \boldsymbol{V}_{n-\deg f+k}\left( \{\boldsymbol{f}_k\} \right) \boldsymbol{h}_f \right)^T \boldsymbol{x}^{n-\deg f+k} = \sum_{k=0}^{\deg f} \left( \boldsymbol{V}_{n-k}\left( \{\boldsymbol{f}_{\deg f-k}\} \right) \boldsymbol{h}_f \right)^T \boldsymbol{x}^{n-k} \\ &= \sum_{k=0}^{n} \left( \boldsymbol{V}_{n-k}\left( \{\boldsymbol{f}_{\deg f-k}\} \right) \boldsymbol{h}_f \right)^T \boldsymbol{x}^{n-k} = \sum_{k=0}^{n} \left( \boldsymbol{V}_k\left( \{\boldsymbol{f}_{\deg f-n+k}\} \right) \boldsymbol{h}_f \right)^T \boldsymbol{x}^k \end{aligned}$$

hence

$$\sum_{f\in F}(h_f f)(x)=\sum_{k=0}^{n}\left(\boldsymbol{V}_k\left(\left\{\boldsymbol{f}_{\deg f-n+k}:f\in F\right\}\right)\boldsymbol{h}\right)^T\boldsymbol{x}^k,$$

and we can compute the reduction of $\boldsymbol{g}$ explicitly as

$$\begin{pmatrix}\boldsymbol{g}_0\\ \vdots\\ \boldsymbol{g}_n\end{pmatrix}-\begin{pmatrix}\boldsymbol{V}_0\left(\left\{\boldsymbol{f}_{\deg f-n}:f\in F\right\}\right)\\ \vdots\\ \boldsymbol{V}_n\left(\left\{\boldsymbol{f}_{\deg f}:f\in F\right\}\right)\end{pmatrix}\boldsymbol{R}_n\Sigma_r^{-1}\boldsymbol{V}_n^H\boldsymbol{g}_n. \tag{2.4.21}$$

The according procedure is given in Algorithm 2.4.2.

**Algorithm 2.4.2** ORTHOGONAL REDUCTION: $F\subset\Pi$ finite set, $g\in\Pi$.

1: $n\leftarrow\deg g$
2: Compute $\boldsymbol{V}_n(F)$ and the SVD (2.4.18)
3: Replace

$$\begin{pmatrix}\boldsymbol{g}_0\\ \vdots\\ \boldsymbol{g}_n\end{pmatrix}\leftarrow\begin{pmatrix}\boldsymbol{g}_0\\ \vdots\\ \boldsymbol{g}_n\end{pmatrix}-\begin{pmatrix}\boldsymbol{V}_0\left(\left\{\boldsymbol{f}_{\deg f-n}:f\in F\right\}\right)\\ \vdots\\ \boldsymbol{V}_n\left(\left\{\boldsymbol{f}_{\deg f}:f\in F\right\}\right)\end{pmatrix}\boldsymbol{R}_n\Sigma_r^{-1}\boldsymbol{V}_n^H\boldsymbol{g}_n$$

**Remark 2.4.27.** Even if the expression in (2.4.21) looks scary, the computational effort is mostly due to the SVD. The $\boldsymbol{V}_k$ matrices are all generated by only shuffling the coefficients of the homogeneous parts and the computation of $\boldsymbol{h}$ is done by means of simple matrix multiplications.

In addition to projection and reduction, we can also find the syzygies of a certain degree in the SVD of the matrix $\boldsymbol{V}_n(F)$.

**Definition 2.4.28.** $s\in S(\lambda(F))$ is called a SYZYGY OF DEGREE $n$ if $s_f\in\Pi^0_{n-\deg f}$.

**Lemma 2.4.29.** *The columns of the matrix $\boldsymbol{S}_n$ from* (2.4.18)*, seen as $s_{f,\alpha}$, $f\in F$, $|\alpha|=n-\deg f$, define a basis of all syzygies of degree $n$ in the sense that*

$$\left(\sum_{|\alpha|=n-\deg f}s_{f,\alpha}x^\alpha:f\in F\right)\in\Pi^F,\qquad\left(s_{f,\alpha}:|\alpha|=n-\deg f,f\in F\right)\in\boldsymbol{S}_n \tag{2.4.22}$$

*are a basis for all syzygies of degree $n$.*

**Proof:** A vector $\boldsymbol{s}=\left(s_{f,\alpha}:|\alpha|=n-\deg f,f\in F\right)$ defines a syzygy as in (2.4.22) if and only $\boldsymbol{V}_n(F)\,\boldsymbol{s}=0$ which in turn holds true if and only if $\boldsymbol{s}=\boldsymbol{S}_n\,\boldsymbol{a}$ for some vector $\boldsymbol{a}$. □

Lemma 2.4.29 can be used to determine syzygies of leading terms and thus perform Buchberger's algorithm: compute the kernels of $\boldsymbol{V}_n(F)$ and reduce any nonzero element in these kernels. The main question, however, is when to stop. This is easy for zero dimensional ideals, cf. [Möller and Sauer, 2000b], but unclear for higher dimensions.

## 2.5 Elimination ideals and intersections

To finish this chapter, we give a very special type of ideals and complete our method for ideal intersection and therefore also for the computation of quotient ideals.

### 2.5.1 Elimination ideals

The idea of elimination ideals is much older than Gröber or even H-bases and dates back to Kronecker. It relies on the "natural" to solve for a simple variable, then eliminate this variable by substituting the solution and thus reducing the number of variables by one. Repeating this, one would end up with all solutions of a system of equations. We will see that lex Gröbner bases play a crucial role here, which was one of the original applications of Gröbner bases and recovered them after almost being forgotten. Indeed, it was a paper by Trinks [Trinks, 1978] that renewed and even triggered the interest in Gröbner bases by focusing on the lex Gröbner basis and its connection to elimination ideals.

**Definition 2.5.1.** For an ideal $\mathscr{I} \subseteq \Pi$, the set

$$\mathscr{I}_k := \mathscr{I} \cap \mathbb{K}[x_1, \ldots, x_k], \qquad k = 1, \ldots, s, \tag{2.5.1}$$

is called the $k$th ELIMINATION IDEAL of $\mathscr{I}$.

Elimination ideals allow us to solve a system of polynomial equations in quite an old-fashioned way:

1. The ideal $\mathscr{I}_1$ is generated by univariate polynomials in $x_1$ only, hence is a principal ideal. We "only" have to compute the gcd of these generators and find the generator for $\mathscr{I}_1$.
2. The zeros of this generator are the $x_1$-components of all common zeros of $\mathscr{I}$.
3. We substitute each of these finitely many zeros for $x_1$ and consider $\mathscr{I}_2$. Since $x_1$ is fixed, this is again a principal ideal, generated by a single polynomial whose zeros we can determine.
4. Proceeding this way, we finally arrive at $\mathscr{I} = \mathscr{I}_n$ and computed all solutions.

So all we need is a basis for the elimination ideals and here the lex term order is exactly the answer.

**Theorem 2.5.2.** *If $\mathscr{I} \subset \Pi$ is an ideal and $G$ a Gröbner basis of $\mathscr{I}$ with respect to the lex term order with $x_1 < x_2 < \cdots < x_n$, then*

$$\mathscr{I}_k = \langle G \cap \mathbb{K}[x_1, \ldots, x_k] \rangle \qquad in\, \mathbb{K}[x_1, \ldots, x_k].$$

**Proof:** We set $G_k = G \cap \mathbb{K}[x_1, \ldots, x_k]$ and fix $k \in 1, \ldots, n$. Any $f \in \mathscr{I}_k \subseteq \mathscr{I}$ has a G-representation

$$f = \sum_{g \in G} f_g\, g, \qquad \delta\left(f_g\, g\right) \leq \delta(f). \tag{2.5.2}$$

If $g \in G \setminus G_k$, then $g$ contains at least one of the variables $x_{k+1}, \ldots, x_n$ which implies that $\delta(g) > \delta(f)$ and the respective coefficient $f_h$ in the G-representation (2.5.2) has to be zero. Therefore,

$$f = \sum_{g \in G_k} f_g\, g,$$

and by the same argument $\delta\left(f_g\right) < \delta(f)$ allows us to conclude that $f_g \in \mathbb{K}[x_1, \ldots, x_k]$. Hence $f$ has a G-representation in $\mathbb{K}[x_1, \ldots, x_k]$ by means of $G_k$ as claimed. □

**Remark 2.5.3.** Unfortunately, the convincing idea of elimination ideals has some severe drawbacks:

1. the lex Gröbner basis is usually the most difficult one to compute and the effort grows *exponentially* in the degree of the polynomials in $F$.
2. The univariate polynomial that generates the principal ideal $\mathscr{I}_1$ and forms the starting point for the computation of the zeros will usually have a very large degree, namely the total number of all common zeros of $\mathscr{I}$. This makes it impossible to explicitly find the zeros of this polynomial and makes numerical computations fairly ill-conditioned.
3. And to make it even worse, the substitution of inaccurate components of the zeros can make the successive computation of other components even worse, and this is in no way a purely theoretical phenomenon.
4. Actually it is a well-known fact, cf. [Farouki and Rajan, 1987, Wilkinson, 1984] that the zeros of a polynomial are very sensitive to changes in the coefficients.

In other words: **Forget it!**

### 2.5.2 Ideal intersection

We still have to finalize the INTERSECTION of two ideals $\mathscr{I}$ and $\mathscr{I}'$, given by bases $F$ and $F'$. By means of (2.1.25) from Lemma 2.1.32 and a proper grading this is now very easy. Indeed, we first compute a basis of $t\mathscr{I} + (1-t)\mathscr{I}'$, namely

$$\left\{tf + (1-t)f' : f \in F, f' \in F'\right\} =: F_\cap \tag{2.5.3}$$

and then find in $H_\cap$ the polynomials that are independent of $t$. To that end, we use a term order with $x \prec t$, for example

$$(x,t)^{\alpha,k} \prec (x,t)^{\alpha',k'} \qquad \Leftrightarrow \qquad \begin{cases} k < k', \\ k = k', \alpha \prec \alpha', \end{cases}$$

and compute a Gröber basis $G \subset \mathbb{K}[x,t]$ of $\langle F_\cap \rangle$ with respect to this term order, and look for all elements in this bases that do not contain $t$, i.e., we consider the elimination ideal

$$\langle F_\cap \rangle \cap \mathbb{K}[x].$$

By Theorem 2.5.2 a basis for this ideal is $G \cap \mathbb{K}[x]$, hence this is also a basis for the intersection. The simple algorithm is summarized in Algorithm 2.5.1.

**Algorithm 2.5.1** IDEAL INTERSECTION: $F, G \subset \Pi$ $\Gamma$-bases.

1: $H \leftarrow \left\{tf(x) + (1-t)g(x) : f \in F, g \in G\right\} \in \mathbb{K}[x,t]$.
2: Compute a $\Gamma$-basis $H$ with $x \prec t$ from $H$
3: **Result:** $\Gamma$-basis $H \cap \mathbb{K}[x]$ for $\langle F \rangle \cap \langle G \rangle$.

# Polynomial zeros | 3

*Any damn fool, he maintained, could think of questions; it was answers that separated the men from the boys. If you couldn't answer your own questions it was either because you hadn't worked on them hard enough or because they weren't real questions.*

(D. Lodge, *The Campus Trilogy*)

Finally, we get to the situation of finding zeros of polynomials. Indeed, we will look at *common* zeros of a finite set of polynomials, or, which is the same, the common zeros of the associated ideal.

## 3.1 Solving equations

**Definition 3.1.1.** Given a finite set $F$, solving the polynomial system of equations $F(X) = 0$ corresponds to finding the zero set

$$Z(F) = \{x \in \mathbb{K}^s : F(x) = 0\} := \{x \in \mathbb{K}^s : f(x) = 0, f \in F\}. \tag{3.1.1}$$

Since for $x \in \mathbb{K}^s$ we almost trivially have that

$$F(x) = 0 \quad \Leftrightarrow \quad \sum_{f \in F} g_f(x)\, f(x) = 0, \quad h_f \in \Pi \quad \Leftrightarrow \quad \langle F \rangle(x) = 0$$

the set $Z(F) = Z(\langle F \rangle)$ depends on the IDEAL $\langle F \rangle$ and not on the specific BASIS $F$.

**Remark 3.1.2.** From an abstract point of view, many algebraic techniques to solve polynomial systems of equations correspond to a CHANGE OF BASIS from which the solutions can be read off more easily. The most prominent case for that is the elimination ideal and the change to a lex Gröbner basis.

### 3.1.1 Zero dimensional ideals and the quotient space

Now we can start to give a formal definition of zero dimensional ideals which will be the only ideals we are interested in in the following. The first concept, based on the preceding chapter is that of the quotient space.

**Definition 3.1.3.** Given an ideal $\mathscr{I}$, a grading $\Gamma$ and an inner product $(\cdot,\cdot)$ on $\Pi \times \Pi$, the NORMAL FORM SPACE or QUOTIENT SPACE or INVERSE SYSTEM $\Pi/\mathscr{I}$ is defined as

$$\Pi/\mathscr{I} = \nu_{\mathscr{I}}(\Pi) = \left\{\nu_{\mathscr{I}}(f) : f \in \Pi\right\}, \tag{3.1.2}$$

where the NORMAL FORM $\nu_{\mathscr{I}}(f)$ is the remainder of division by the division with remainder of Algorithm 2.3.2 with respect to a $\Gamma$-basis of $\mathscr{I}$.

**Remark 3.1.4** (Quotient spaces)**.**

1. By 2.3.30 the normal form $\nu_{\mathscr{I}}$ is independent of the $\Gamma$-basis, hence $\Pi/\mathscr{I}$ is indeed well-defined and depends on $\mathscr{I}$ only.

2. The linearity of the reduction process implies that

$$\nu_{\mathscr{I}}(f+g) = \nu_{\mathscr{I}}(f) + \nu_{\mathscr{I}}(g), \qquad \nu_{\mathscr{I}}(c\,f) = c\,\nu_{\mathscr{I}}(f), \qquad f,g \in \Pi,\, c \in \mathbb{K},$$

   hence $\Pi/\mathscr{I}$ is a $\mathbb{K}$-vector space.

3. The normal form is an IDEAL PROJECTION:

$$\nu_{\mathscr{I}}\left(\nu_{\mathscr{I}}(f)\right) = \nu_{\mathscr{I}}(f) \tag{3.1.3}$$

   and

$$\nu_{\mathscr{I}}(f) = 0 \qquad \Leftrightarrow \qquad f \in \mathscr{I}. \tag{3.1.4}$$

4. The name "inverse system" is due to Macaulay and has been followed up by Gröbner in [Gröbner, 1937]. In the world of Gröbner bases, the name has almost been forgotten, however.

**Example 3.1.5.** On standard example for the homogeneous grading is that $\mathscr{I} = \langle F \rangle$ where $F \subset \Pi_n$ and $\Pi_n^0 = \operatorname{span} \lambda(F)$. Then

$$V_k(F) = \begin{cases} \{0\}, & k < n, \\ \Pi_k^0, & k \geq n, \end{cases} \qquad W_k(F) = \begin{cases} \Pi_k^0, & k < n, \\ \{0\}, & k \geq n, \end{cases} \qquad k \in \mathbb{N}_0.$$

Then $\Pi/\mathscr{I} = W(F) = \Pi_{n-1}$ gives the total degree analogy of the univariate case. Note that under the condition that $\Pi_n^0 = \operatorname{span} \lambda(F)$ the quotient space depends only the total degree $n$.

**Example 3.1.6.** As a more concrete and very simple situation we consider the following three bases in $\mathbb{C}[x,y]$, namely

$$F_1 = \left\{x(x-1), xy, y(y-1)\right\}, \qquad F_2 = \left\{x(x-1), xy, y^2\right\}, \qquad F_3 = \left\{x^2, xy, y^2\right\}. \tag{3.1.5}$$

In all three cases $\lambda(F_j) = \{x^2, xy, y^2\}$, $\Pi/\langle \mathscr{I}_j \rangle = \Pi_1$, $\mathscr{I}_j := \langle F_j \rangle$, while is can be easily seen that the zero sets $Z_j := Z(\mathscr{I}_J)$ satisfy

$$Z_1 = \{(0,0),(1,0),(0,1)\}, \qquad Z_2 = \{(0,0),(1,0)\}, \qquad Z_3 = \{(0,0)\}, \tag{3.1.6}$$

In other words, the quotient space alone does not give information about the underlying ideal.

**Definition 3.1.7.** The ideal $\mathscr{I}$ is called ZERO DIMENSIONAL if $\dim \Pi/\mathscr{I} < \infty$.

Next we will show that any zero dimensional ideal has only finitely many common zeros by extending the multiplication operators from (1.1.22) in an almost straightforward way. However, we will have to slightly vary the notation.

**Definition 3.1.8.** For $q \in \Pi$ the MULTIPLICATION OPERATOR $M_q : \Pi/\mathscr{I} \to \Pi/\mathscr{I}$ is defined as

$$M_q f := M[q] f := \nu_{\mathscr{I}}(q\,f), \qquad f \in \Pi/\mathscr{I}. \tag{3.1.7}$$

**Lemma 3.1.9.** *The multiplication operators are linear and commute, i.e., for* $f, f' \in \Pi/\mathscr{I}$, $c, c' \in \mathbb{K}$ *and* $q, q' \in \Pi$ *we have*

$$M_q(cf + c'f') = c\,M_q f + c'\,M_q f', \qquad M_q M_{q'} f = M_{qq'}(f) = M_{q'} M_q f. \tag{3.1.8}$$

**Proof:** The first property follows from the linearity of normal forms,

$$M_q(cf + c'f') = \nu_{\mathscr{I}}\left(q(cf + c'f')\right) = c\,\nu_{\mathscr{I}}\left(qf\right) + c'\,\nu_{\mathscr{I}}\left(qf'\right),$$

for the second one we write $q'f = \nu_{\mathscr{I}}(q'f) + g$, $g \in \mathscr{I}$, and note that

$$\begin{aligned} M_{qq'}f &= \nu_{\mathscr{I}}\left(qq'f\right) = \nu_{\mathscr{I}}\left(q\left(\nu_{\mathscr{I}}(q'f) + g\right)\right) = \nu_{\mathscr{I}}\left(q\,\nu_{\mathscr{I}}(q'f) + qg\right) \\ &= \nu_{\mathscr{I}}\left(q\,\nu_{\mathscr{I}}(q'f)\right) + \nu_{\mathscr{I}}\left(qg\right) = M_q M_{q'} f, \end{aligned}$$

since $qg \in \mathscr{I}$ implies that $\nu_{\mathscr{I}}\left(qg\right) = 0$. □

**Remark 3.1.10.** Note that the proof of the commuting property only uses the fact that the normal form is a linear operator whose kernel is an ideal.

Next, we again choose a basis $P$ for the finite dimensional vector space $\Pi/\mathscr{I}$ and represent the multiplication operator in terms of $P$,

$$M_q := M[q] = \left(m^q_{p,p'} : p, p' \in P\right) \in \mathbb{K}^{P\times P}, \qquad M_q p = \sum_{p'\in P} m^q_{p,p'}\, p'. \tag{3.1.9}$$

By Lemma 3.1.9, any matrices $M_q$ and $M_{q'}$, $q, q' \in \Pi$ commute.

**Definition 3.1.11.** The matrix $M_q$ from (3.1.9) is called to MULTIPLICATION TABLE for multiplication by $q$ with respect to the basis $P$.

**Remark 3.1.12.** From now on, we use $M_q$ to denote the operator and its matrix representation with respect to a basis $P$. To keep notation simple, we will not specify the basis $P$ explicitly, even if it can significantly affect the matrices as shown in Example 1.1.22 for the univariate case. Usually, we will always start with an ideal that defines a quotient space and then assume that a basis for this quotient space has been fixed.

**Example 3.1.13.** Let us compute the multiplication tables $M_x$ and $M_y$ for the three ideals.

1. With respect to $\mathscr{I}_1$ we get the multiplications

$$x\cdot 1 = x \to x, \qquad x\cdot x = x^2 \to x, \qquad x\cdot y = xy \to 0 \qquad \Rightarrow \qquad M_x = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

   and analogously

$$y\cdot 1 = y \to y, \qquad y\cdot x = xy \to 0, \qquad y\cdot y = y^2 \to y \qquad \Rightarrow \qquad M_y = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{pmatrix}.$$

   Here rows and columns are indexed by means of the basis elements $(1, x, y)$ of $\Pi_1$.

2. In the same way, we obtain for $\mathscr{I}_2$ that

$$x\cdot 1 = x \to x, \qquad x\cdot x = x^2 \to x, \qquad x\cdot y = xy \to 0 \qquad \Rightarrow \qquad M_x = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

and analogously

$$y\cdot 1 = y \to y, \qquad y\cdot x = xy \to 0, \qquad y\cdot y = y^2 \to 0 \qquad \Rightarrow \qquad M_y = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

3. The simplest case, however, is $\mathscr{I}_3$ where we have that

$$x\cdot 1 = x \to x, \qquad x\cdot x = x^2 \to 0, \qquad x\cdot y = xy \to 0 \qquad \Rightarrow \qquad M_x = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

and analogously

$$y\cdot 1 = y \to y, \qquad y\cdot x = xy \to 0, \qquad y\cdot y = y^2 \to 0 \qquad \Rightarrow \qquad M_y = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

Now the companion matrix idea extends to the multivariate case using multiplication tables.

**Theorem 3.1.14.** *For $z \in Z(\mathscr{I})$ and $q \in \Pi$, the number $q(z)$ is a eigenvalue of $M_q$ and the associated eigenvector is independent of $q$ if $\mathscr{I}$ is radical.*

**Lemma 3.1.15.** *A zero dimensional ideal has only finitely many zeros, more precisely*

$$\#Z(\mathscr{I}) \le \dim \Pi/\mathscr{I}. \tag{3.1.10}$$

**Proof:** Let $Z \subseteq Z(\mathscr{I})$ be any finite set of common zeros of $\mathscr{I}$ and let, for given $z \in Z$, $v_{z,z'} \in \mathbb{K}^s$, $z' \in Z \setminus \{z\}$ be vectors such that $v_{z,z'}^T(z-z') \neq 0$, $z' \in Z \setminus \{z\}$. We define the polynomials

$$\ell_z := \nu_{\mathscr{I}}\left(\prod_{z'\in Z} \frac{v_{z,z'}^T\left(\cdot - z'\right)}{v_{z,z'}^T\left(z-z'\right)}\right) \in \Pi/\mathscr{I}, \qquad z \in Z. \tag{3.1.11}$$

Since

$$\ell_z := \prod_{z'\in Z} \frac{v_{z,z'}^T\left(\cdot - z'\right)}{v_{z,z'}^T\left(z-z'\right)} + g$$

for some $g \in \mathscr{I}$, it follows that

$$\ell_z(z') = \delta_{z,z'}, \qquad z, z' \in Z.$$

and in particular the polynomials $\ell_z$, $z \in Z$, are linearly independent which implies that $\#Z \le \dim \Pi/\mathscr{I}$. Since $Z$ was an arbitrary finite subset of $Z(\mathscr{I})$, this yields (3.1.10). □

This already gives a first view in the direction of interpolation.

**Corollary 3.1.16.** *If $\mathscr{I}$ is zero dimensional, the space $\Pi/\mathscr{I}$ always allows for interpolation at $Z(\mathscr{I})$ by means of*

$$Lf := \sum_{z\in Z(\mathscr{I})} f(z)\,\ell_z, \tag{3.1.12}$$

*which satisfies $L_n f(z) = f(z)$, $z \in Z(\mathscr{I})$.*

**Proof of Theorem 3.1.14:** By Lemma 3.1.15, the set $Z(\mathscr{I})$ is finite and we can again use the finitely many polynomials $\ell_z$, $z \in Z(\mathscr{I})$, defined in (3.1.11). Then, for $z \in Z(\mathscr{I})$

$$M_q\ell_z = \nu_{\mathscr{I}}(q\ell_z) = q\ell_z + g = \sum_{z'\in Z} q(z')\ell_z(z')\ell_{z'} + g_z + g = q(z)\ell_z + g_z + g \tag{3.1.13}$$

for some $g \in \mathscr{I}$, $g_z \in \sqrt{\mathscr{I}}$. If $\mathscr{I}$ is radical, then indeed $M_q\ell_z = q(z)\ell_z$, otherwise the subspace

$$L^\perp := \{f \in \Pi/\mathscr{I} : Lf = 0\}$$

is nontrivial and we have the decomposition

$$\Pi/\mathscr{I} = L(\Pi/\mathscr{I}) + L^\perp.$$

For $g \in L^\perp$ we have that

$$M_q g = \nu_{\mathscr{I}}(qg) = \sum_{z'\in Z} q(z')g(z')\ell_{z'} + g' = g' \in L^\perp,$$

hence $L^\perp$ is invariant under $M_q$ for any $q$. Assuming that $L_z^\perp \setminus \{0\} \ni g_z := M_q\ell_z - q(z)\ell_z$, as otherwise $\ell_z$ is an eigenvector, we consider the KRYLOV SPACES

$$K_n := \operatorname{span}\left\{\left(M_q - q(z)I\right)^k g_z : k = 0,\dots,n-1\right\}, \qquad k \ge 1,$$

and note that[1] $\dim K_n = \max\{n, n_0\}$ for some $n_0 \ge 1$. Hence, there exists some sequence $\alpha_1,\dots,\alpha_{n_0}$ such that

$$g_z = \sum_{k=1}^{n_0} \alpha_k \left(M_q - q(z)I\right)^k g_z.$$

Setting

$$g_z^* := \sum_{k=1}^{n_0} \alpha_k \left(M_q - q(z)I\right)^{k-1} g_z$$

we then get that

$$\begin{aligned}\left(M_q - q(z)I\right)\left(\ell_z - g_z^*\right) &= g_z - \left(M_q - q(z)I\right)\left(\sum_{k=1}^{n_0} \alpha_k \left(M_q - q(z)I\right)^{k-1} g_z\right)\\ &= g_z - \sum_{k=1}^{n_0} \alpha_k \left(M_q - q(z)I\right)^k g_z = 0,\end{aligned}$$

hence $\ell_z - g_z^*$ is an eigenvector for the eigenvalue $q(z)$ as claimed. □

[1] Indeed, if $K_{n+1} = K_n$, that is, $\left(M_q - q(z)I\right)^n g_z \in K_n$, we can conclude that

$$\left(M_q - q(z)I\right)^n g_z = \sum_{k=0}^{n-1} \alpha_k \left(M_q - q(z)I\right)^n g_z$$

hence,

$$\left(M_q - q(z)I\right)^{n+1} g_z = \left(M_q - q(z)I\right)\left(\sum_{k=0}^{n-1} \alpha_k \left(M_q - q(z)I\right)^n g_z\right) \in K_{n+1} = K_n,$$

so that the dimension either increases by one or stays constant forever. This is a standard argument from numerical linear algebra and useful, for example, in the context of CONJUGATE GRADIENTS.

**Remark 3.1.17.** If the zero dimensional ideal $\mathscr{I}$ is a radical ideal then $\dim \Pi/\mathscr{I} = \#Z(\mathscr{I})$ and therefore the polynomials $\ell_z$, $z \in Z(\mathscr{I})$ form a basis of the quotient space. Therefore, $L^\perp = \{0\}$ is trivial and the $\ell_z$ as defined in (3.1.11) are the *unique* elements of the quotient space with $\ell_z(z') = \delta_{z,z'}$, $z, z' \in Z(\mathscr{I})$. Otherwise $\ell_z$ is only defined up to an element from $L^\perp$, i.e., $\ell'_z = \ell_z + g_z$, $g_z \in L^\perp$ also have the same interpolation properties, and this ambiguity is reflected in the fact that we do not get eigenvectors but only principal vectors. And it is not surprise that the above proof was using concepts of the proof of the Jordan normal form in [Fischer, 1984].

**Example 3.1.18.** We can determine the eigenstructure, more precisely the JORDAN NORMAL FORM of the multiplication tables in each of our three examples. The normal form and the generalized eigenvalues are computed by the `symbolic` toolbox of `octave`. The call looks, for example, as follows[2]

```
>> pkg load symbolic;
>> A = sym( [ 0 1 0 ; 0 1 0 ; 0 0 0 ] );
>> [gevs,JnF] = jordan( A)
gevs = (sym 3×3 matrix)

  1  0  1
  0  0  1
  0  1  0

JnF = (sym 3×3 matrix)

  0  0  0
  0  0  0
  0  0  1
```

and gives a matrix of generalized eigenvalues and the Jordan normal form of the matrix. The Jordan normal forms are now as follows:

$$\begin{aligned}
\mathscr{I}_1 \qquad & M_x \sim \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad M_y \sim \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \\
\mathscr{I}_2 \qquad & M_x \sim \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad M_y \sim \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \\
\mathscr{I}_3 \qquad & M_x \sim \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad M_y \sim \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.
\end{aligned}$$

This shows the difference in the multiplication tables: in the radical ideal $\mathscr{I}_1$ both multiplication tables are diagonalizable as predicted by Theorem 3.1.14, for $\mathscr{I}_2$ one multiplication table is diagonalizable, the other one, $M_y$, has a nontrivial Jordan block and $\mathscr{I}_3$ has Jordan blocks in both multiplication tables.

Already the simple Example 3.1.18 shows that we have to be a bit more careful with the eigenstructure of the multiplication tables that seems to tell us a lot about the underlying structure.

[2] After installing the package from the *Sourceforge* page and calling `pgk install` from `Octave`.

### 3.1.2 Making ideals radical

Theorem 3.1.14 shows that the situation is favorable if the underlying $\mathscr{I}$ is a radical ideal. This corresponds to the univariate case where we assumed the zeros to be simple. In one variable it is easy to get rid of multiplicities by simply considering $f/\gcd(f,f')$ which has the same zeros as $f$ but without multiplicities and is easy to compute. In several variables this will also be possible and give us a tool to compute the radical of a given ideal. And it will be based on multiplication tables again.

**Remark 3.1.19.** Even in a single variable, multiple zeros are troublemakers, be it for algorithms like Newton's method [Gautschi, 1997] or for eigenvalue methods. Therefore, removing multiplicities is always worthwhile if one is interested only in finding the variety $Z(\mathscr{I})$.

In terms of computational algebra our goal is, given a basis $F$ for the ideal $\langle F\rangle$ to determine the a basis $G$ for the radical

$$\langle G\rangle = \sqrt{\langle F\rangle} := \left\{ g \in \Pi \,:\, g^k \in \langle F\rangle,\, k \geq k_0 \right\}$$

whose zeros can then be determined by the eigenvalue method – even if this will need some more effort. To that end we use the so-called TRACE METHOD from [González-Vega et al., 1999]. Again, we shift the focus from the ideal to the computationally accessible basis of the basis.

To achieve the goal of computing the radical, we need a refined version of Theorem 3.1.14 in the form of a generalized eigenspaces decomposition based on a replacement of the functions $\ell_z$. For this approach that is a modification of the one in [Cohen et al., 1999, Chapter 2] we need an algebraically closed field now.

**Proposition 3.1.20.** *Let $\mathscr{I} \subset \mathbb{C}[x]$ be a zero dimensional ideal. Then there exist polynomials $m_z$, $z \in Z(\mathscr{I})$, such that*

$$m_z^2 = m_z, \qquad m_z(z') = \delta_{z,z'}, \qquad m_z m_{z'} = 0, \qquad z, z' \in Z(\mathscr{I}), \tag{3.1.14}$$

*and*

$$\sum_{z \in Z(\mathscr{I})} m_z = 1. \tag{3.1.15}$$

**Remark 3.1.21.** The properties (3.1.14) and (3.1.14) are satisfied by the $\ell_z$ from the preceding chapter if $\mathscr{I}$ is a radical ideal, in general we can only ensure that

$$1 - \sum_z \ell_z,\ \ell_z^2 - \ell_z,\ \ell_z \ell_{z'} \in L^\perp$$

which would be too weak for what follows.

**Proof:** We nevertheless begin with the $\ell_z$ and note that $(\ell_z \ell_{z'})(Z(\mathscr{I})) = 0$ yields, by means of the Nullstellensatz, Theorem 2.1.16, that $(\ell_z \ell_{z'})^m \in \mathscr{I}$ for some $m \geq 1$, hence, $p_z := \ell_z^m \in \Pi/\mathscr{I}$ satisfies[3] $p_z\, p_{z'} = 0$. Moreover,

$$p_z(z) = \ell_z(z)^m = 1^m = 1$$

Therefore, the polynomials in $\mathscr{I}$ and $\{p_z : z \in Z(\mathscr{I})\}$ have no common zeros in the algebraically closed field $\mathbb{C}$, so that

$$\left\langle \mathscr{I} \cup \{p_z : z \in Z(\mathscr{I})\} \right\rangle = \Pi \qquad \Rightarrow \qquad 1 = g + \sum_{z \in Z(\mathscr{I})} q_z\, p_z \quad \text{for some } g \in \mathscr{I},\, q_z \in \Pi,$$

[3] Since we work modulo $\mathscr{I}$ here.

and the polynomials $m_z := \nu_{\mathscr{I}}\left(q_z p_z\right)$, $z \in Z(\mathscr{I})$, satisfy (3.1.15). In addition,

$$m_z(z') = q_z(z')\ell_z(z')^m = 0, \qquad z' \neq z,$$

yields, together with (3.1.15), that $m_z(z) = 1$, while $m_z(z') = 0$, $z' \neq z$, follows from the fact that $m_z$ is a multiple of $\ell_z$. Finally,

$$m_z = m_z 1 = m_z\left(\sum_{z\in Z(\mathscr{I})} m_z\right) = m_z^2 + \sum_{z'\neq z} m_z m_{z'} = m_z^2$$

completes the last claim in (3.1.14). □

**Remark 3.1.22.** It has to be emphasized that all multiplications in the above proof are performed in the quotient space $\Pi/\mathscr{I}$, hence modulo $\mathscr{I}$. Note however, that this is **not** an integral ring, there are lots of zerodivisors.

The polynomials $m_z$ help us to decompose $\Pi/\mathscr{I}$ into principal subspaces with good properties.

**Proposition 3.1.23.** *The subspaces*

$$\mathscr{M}_z := m_z \cdot (\Pi/\mathscr{I}) = \left\{\nu_{\mathscr{I}}(m_z f) : f \in \Pi/\mathscr{I}\right\}, \qquad z \in Z(\mathscr{I}), \tag{3.1.16}$$

*form a direct sum decomposition of* $\Pi/\mathscr{I}$ *and are invariant under* any *multiplication operator with eigenvalue* $q(z)$.

**Proof:** Again, linearity of the remainder[4] implies that any $\mathscr{M}_z$ is a linear subspace of $\Pi/\mathscr{I}$. That any $f \in \Pi/\mathscr{I}$ can be represented as a sum follows from

$$f = f \cdot 1 = f\left(\sum_{z\in Z(\mathscr{I})} m_z\right) = \sum_{z\in Z(\mathscr{I})} f\, m_z$$

while uniqueness is a consequence of

$$0 = \sum_{z\in Z(\mathscr{I})} f_z\, m_z \qquad \Rightarrow \qquad 0 = m_z\left(\sum_{z'\in Z(\mathscr{I})} f_{z'}\, m_{z'}\right) = f_z m_z^2 = f_z m_z,$$

hence all components of the representation must be zero. Invariance under multiplication follows directly from the definition of $\mathscr{M}_z$. Finally, we consider, for $q \in \Pi$

$$q\, m_z = \big(q - q(z)\big)\, m_z + q(z)\, m_z$$

and note that since $\big(q - q(z)\big)\, m_z$ vanishes at $Z(\mathscr{I})$, there exists some $n \geq 1$ such that $0 = \big(q - q(z)\big)^n\, m_z$. Hence, by the first property of (3.1.14)

$$0 = (M_q - q(z)\, I)^n\, m_z = (M_q - q(z)\, I)^n\, m_z^n = \big((M_q - q(z)\, I)\, m_z\big)^n$$

implies that $(M_q - q(z)\, I)\, m_z$ is a nilpotent element in $\mathscr{M}_z$ and therefore $q(z)$ is the unique eigenvalue associated to this space; the eigenvector can be constructed by from $m_z$ by the Krylov space method used in the proof of Theorem 3.1.14. □

[4] Just recall that this follows from the uniqueness of remainder when dividing by a $\Gamma$-basis as then both $\nu_{\mathscr{I}}(f) + \nu_{\mathscr{I}}(g)$ and $\nu_{\mathscr{I}}(f+g)$ are candidates for the same remainder, hence must coincide.

**Definition 3.1.24.** Let $P$ be a basis $\Pi/\langle F\rangle$, $\langle F\rangle$ zero dimensional.

1. For $z \in Z(\mathscr{I})$ denote by $\mu(z)$ the MULTIPLICITY of the zero, defined as
$$\mu(z) = \dim \mathscr{M}_z. \tag{3.1.17}$$
2. For $q \in \Pi$ we define the TRACE MATRIX $T_q \in \mathbb{K}^{P\times P}$ as
$$T_q = \left(\operatorname{trace} M\left[qpp'\right] \,:\, p,p' \in P\right). \tag{3.1.18}$$

**Remark 3.1.25.** Defining multiplicity of a zero as in (3.1.17), that is, as a number, is insufficient and ignores structural concepts. We will give a more appropriate and powerful definition soon.

**Example 3.1.26.** That counting is insufficient, can already be seen by the following simple example: A simple common zero $z$ means that $f(z) = 0$, a double zero that $f(z) = 0$ and that some DIRECTIONAL DERIVATIVE $q(D)f(z) = 0$, $q \in \Pi_1^0$, but a TRIPLE ZERO could either involve a second, linearly independent, directional derivative $q'(D)f(z) = 0$, $q' \in \Pi_1^0$, or a second order derivative, in that case $q^2(D)f(z) = 0$, with the $q$ from the double zero.

**Exercise 3.1.1** Show: If $\#F = 1$, then any zero has multiplicity at least $s$. ♢

**Example 3.1.27.** To derive the trace matrices for $\mathscr{I}_1$, $\mathscr{I}_2$ and $\mathscr{I}_3$, we have to compute $M_1 = 1$, $M_x$, $M_y$ and $M_{xy}$, $M_{x^2}$ and $M_{y^2}$. To compute $M_{xy}$ we note that for $\mathscr{I}_1$ we get

$$xy\cdot 1 = xy \to 0, \qquad xy\cdot x = x^2y \to 0, \qquad xy\cdot y = xy^2 \to 0 \qquad \Rightarrow \qquad M_{xy} = 0,$$

but since we only used the common basis element $xy$ for reduction, the same holds true for the other two ideals as well.

1. For $\mathscr{I}_2$ the remaining two multiplication tables compute as
$$x^2\cdot 1 \to x, \qquad x^2\cdot x \to x^2 \to x, \qquad x^2\cdot y \to 0 \qquad \Rightarrow \qquad M_{x^2} = \begin{pmatrix} 0 & 0 & 0\\ 1 & 1 & 0\\ 0 & 0 & 0\end{pmatrix},$$
and
$$y^2\cdot 1 \to y, \qquad y^2\cdot x \to 0, \qquad y^2\cdot y \to y^2 \to y \qquad \Rightarrow \qquad M_{y^2} = \begin{pmatrix} 0 & 0 & 0\\ 0 & 0 & 0\\ 1 & 0 & 1\end{pmatrix},$$
2. for $\mathscr{I}_2$ we get
$$x^2\cdot 1 \to x, \qquad x^2\cdot x \to x^2 \to x, \qquad x^2\cdot y \to 0 \qquad \Rightarrow \qquad M_{x^2} = \begin{pmatrix} 0 & 0 & 0\\ 1 & 1 & 0\\ 0 & 0 & 0\end{pmatrix},$$
as well as
$$y^2\cdot 1 \to 0, \qquad y^2\cdot x \to 0, \qquad y^2\cdot y \to 0 \qquad \Rightarrow \qquad M_{y^2} = 0$$

3. while $\mathscr{I}_3$ gives

$$x^2 \cdot 1 \to 0, \qquad x^2 \cdot x \to 0 \qquad x^2 \cdot y \to 0 \qquad \Rightarrow \qquad M_{x^2} = 0$$

and

$$y^2 \cdot 1 \to 0, \qquad y^2 \cdot x \to 0, \qquad y^2 \cdot y \to 0 \qquad \Rightarrow \qquad M_{y^2} = 0.$$

Thus, the symmetric trace matrices

$$T = T_1 : \begin{pmatrix} \operatorname{trace} M_1 & \operatorname{trace} M_x & \operatorname{trace} M_y \\ \operatorname{trace} M_x & \operatorname{trace} M_{x^2} & \operatorname{trace} M_{xy} \\ \operatorname{trace} M_y & \operatorname{trace} M_{xy} & \operatorname{trace} M_{y^2} \end{pmatrix}$$

take the form

$$T_1(\mathscr{I}_1) = \begin{pmatrix} 3 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}, \qquad T_1(\mathscr{I}_2) = \begin{pmatrix} 3 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad T_1(\mathscr{I}_3) = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

and the kernels are trivial or spanned by the third or the second and the third unit vector, respectively.

The symmetric trace matrix is very useful. To see why, we consider a coefficient vector $v = \left(v_p : p \in P\right) \in \mathbb{C}^p$ and the associated polynomial $f = v \cdot P \in \Pi/\mathscr{I}$. Since the trace of a matrix is the sum of its eigenvalues and since the eigenvalues of $M_q$ are $q(z)$, $z \in Z(\mathscr{I})$ with multiplicity $\mu(z)$, we obtain for $v = \left(v_p : p \in P\right) \in \mathbb{C}^P$ that

$$\begin{aligned} v^H T_q\, v &= \sum_{p,p' \in P} \left(T_q\right)_{p,p'} \overline{v_p}\, v_{p'} = \sum_{p,p' \in P} \left(\operatorname{trace} M\left[q p p'\right]\right) \overline{v_p}\, v_{p'} \\ &= \operatorname{trace} M\left[\sum_{p,p' \in P} q \cdot \overline{v_p}\, p \cdot v_{p'}\, p'\right] = \operatorname{trace} M\left[q \cdot |f|^2\right], \qquad f = v \cdot P = \sum_{p \in P} v_p\, p \in P, \\ &= \sum_{z \in Z(\mathscr{I})} \mu(z)\, q(z)\, |f(z)|^2, \end{aligned}$$

hence

$$v^H T_q\, v = \sum_{z \in Z(\mathscr{I})} \mu(z)\, q(z)\, |(v \cdot P)(z)|^2. \tag{3.1.19}$$

This already suffices to prove a statement that characterizes the polynomials in $\Pi/\mathscr{I}$ which vanish on $Z(\mathscr{I})$, that is, the polynomials that prevent $\mathscr{I}$ from being radical.

**Theorem 3.1.28.** *For a polynomial $f = v \cdot P \in \Pi/\mathscr{I}$, $v \in \mathbb{C}^P$, we have that*

$$f(Z(\mathscr{I})) = 0 \qquad \Leftrightarrow \qquad T_1\, v = 0. \tag{3.1.20}$$

**Proof:** "⇒": for $w \in \mathbb{C}^n$ we define $g := w \cdot P$ and get that

$$w^T T_1\, v = \sum_{z \in Z(\mathscr{I})} \mu(z)\, g(z) \underbrace{f(z)}_{=0} = 0,$$

and since $w$ was arbitrary, this implies that $T_1\, v = 0$.

**Algorithm 3.1.1** RADICAL COMPUTATION: $F \subset \Pi$ such that $\langle F \rangle$ is zero dimensional.

1: Compute a $\Gamma$-basis $G$ of $\langle F \rangle$ and the quotient space $\Pi/\langle G \rangle = \nu_G(\Pi)$.
2: Compute a basis $P$ of $\Pi/\langle G \rangle$.
3: Compute the multiplication tables $M_{pp'}$, $p, p' \in P$ and the trace matrix

$$T_1 \leftarrow \left(\operatorname{trace} M_{pp'} : p, p' \in P\right)$$

4: Determine by SVD a basis $V$ of

$$\ker T_1 := \left\{ v \in \mathbb{K}^{|P|} \,:\, T_1\, v = 0 \right\}.$$

5: Compute a $\Gamma$-basis $R$ of $\langle F \cup V \cdot P \rangle$
6: **Result:** $\Gamma$-basis $R$ such that $\sqrt{\mathscr{I}} = \langle R \rangle$.

For "$\Leftarrow$" assume that $T_1\, v = 0$ and let $v_z \in \mathbb{C}^P$ be the linear independent vectors such that $m_z = v_z \cdot P$, $z \in Z(\mathscr{I})$, with the $m_z$ from (3.1.14). Then we get that

$$0 = v_z^T\, 0 = v_z^T\, T_1\, v = \sum_{z' \in Z(\mathscr{I})} \mu\left(z'\right) \underbrace{m_z\left(z'\right)}_{=\delta_{z,z'}} f\left(z'\right) = f(z), \qquad z \in Z(\mathscr{I}),$$

from which we conclude that $f\left(Z(\mathscr{I})\right) = 0$. □

This eventually allows us to design an algorithm that removes multiplicities of zeros or, in other words, computes the radical of an ideal.

**Example 3.1.29.** We can now give the radicals of our examples, namely $\sqrt{\mathscr{I}_1} = \mathscr{I}_1$

$$\sqrt{\mathscr{I}_2} = \left\langle x(x-1), xy, y^2, y \right\rangle = \left\langle x(x-1), y \right\rangle, \qquad \sqrt{\mathscr{I}_3} = \left\langle x^2, xy, y^2, x, y \right\rangle = \left\langle x, y \right\rangle,$$

which are exactly the ideals of all polynomials vanishing at $(0,0)$ and $(1,0)$ or $(0,0)$, respectively.

### 3.1.3 Finding the zeros

Now we only have to put together the tools we built so far to get an algorithm for finding the common zeros of a zero dimensional ideal. After passing to the radical, the multiplication tables are diagonalizable and have only Jordan blocks of size $1 \times 1$ and since the respective eigenvectors are the $\ell_z$, $z \in Z\left(\sqrt{\mathscr{I}}\right)$, we can use them to connect the solution. Especially the eigenvalues of the *coordinate multiplications* $M_{(\cdot)_j}$, $j = 1, \ldots, s$, are the coordinate projections $z_j$ of the zeros.

**Remark 3.1.30.** In principle, Algorithm 3.1.2 works well, but it runs into problems if there are $z, z' \in Z(\mathscr{I})$ and $j \in \{1, \ldots, s\}$ such that $z_j \approx z'_j$ as then the eigenvalue of $M_{(\cdot)_j}$ becomes a double one and the eigenvectors cannot be determined any more – any linear combination of the two eigenvectors is an eigenvector again.

To overcome the problem of multiple eigenvalues in the multiplication tables one could switch to different coordinate projections as any set of multiplication tables $M_{v_j^T(\cdot)}$ with linearly independent $v_j \in \mathbb{C}^s$ would do the job; however, predicting the proper $v_j$ is difficult again, though they are "bad" only on a set of measure zero. We will use a different approach that directly computes the joint eigenvalues of commuting matrices, thus making use of the structure of the multiplication tables. This approach will be derived in Section 3.1.4.

**Algorithm 3.1.2** COMMON ZEROS: $F \subset \Pi$ such that $\langle F \rangle$ is zero dimensional.

1: Compute a $\Gamma$-basis $G$ of $\sqrt{\langle F \rangle}$ by Algorithm 3.1.1.
2: In $\Pi/\langle G \rangle$ compute the multiplication tables $M_{(\cdot)_j}$.
3: Compute the eigenvalues $\lambda_{jk}$ and eigenvectors and $v_{jk}$ of $M_{(\cdot)_j}$, $k = 1,\ldots,\dim \Pi/\langle G \rangle$.
4: **for do** $k = 1,\ldots,\dim \Pi/\langle G \rangle$
5: $\quad v \leftarrow v_{1k}$
6: $\quad z^k \leftarrow \left(\lambda_{jk'} : v_{jk'} = v,\ j = 1,\ldots,s\right)$
7: **end for**
8: **Result:** Common zeros $z^k$, $k = 1,\ldots,\dim \Pi/\langle G \rangle$

### 3.1.4 Common eigenvectors of commuting families of matrices

Before we define the algorithm for computing the common eigenvectors, we first give some special ones among them a special name.

**Definition 3.1.31.** The $j$th COMPANION MATRIX $M_j$ is defined as the multiplication table $M_{(\cdot)_j}$, $j = 1,\ldots,s$.

**Remark 3.1.32.** According to Theorem 3.1.9, the companion matrices form a commuting family of matrices and these matrices are DIAGONALIZABLE if the ideal $\mathscr{I}$ is a radical ideal as in this case the polynomial $\ell_z$, $z \in Z(\mathscr{I})$, form an eigenvector basis of $\Pi/\mathscr{I}$. Writing $\ell_z = v_z \cdot P$, $v_z \in \mathbb{C}^P$, for some basis $P$ of the quotient space, and setting $V = [v_z : z \in Z(\mathscr{I})] \in \mathbb{C}^{P \times Z(\mathscr{I})}$, it follows that

$$M_j V = V \operatorname{diag}\left(z_j : z \in Z(\mathscr{I})\right) \quad \Leftrightarrow \quad V^{-1} M_j V = \operatorname{diag}\left(z_j : z \in Z(\mathscr{I})\right), \quad j = 1,\ldots,s, \tag{3.1.21}$$

hence the companion matrices for a radical ideal are JOINTLY DIAGONALIZABLE by means of $V$.

**Example 3.1.33.** The case $\mathscr{I}_1$ of Example 3.1.6 shows which problem can nevertheless occur. If we use eigenvalue computations of $M_x$ and $M_y$ be means of `octave`, we get

```
>> [l,E] = eig( [ 0 0 0 ; 1 1 0 ; 0 0 0 ] )
l =

   0.00000   0.70711   0.00000
   1.00000  -0.70711   0.00000
   0.00000   0.00000   1.00000

E =

Diagonal Matrix

   1   0   0
   0   0   0
   0   0   0
```

and

```
>> [l,E] = eig( [ 0 0 0 ; 0 0 0 ; 1 0 1 ] )
l =

   0.00000   0.70711   0.00000
   0.00000   0.00000   1.00000
   1.00000  -0.70711   0.00000

E =

Diagonal Matrix

   1   0   0
   0   0   0
   0   0   0
```

which allows us to identify the zeros $(1,0)$ and $(0,1)$ by means of the eigenvectors $\begin{pmatrix}0\\0\\1\end{pmatrix}$ and $\begin{pmatrix}0\\1\\0\end{pmatrix}$.

The third common eigenvector, $\begin{pmatrix}1\\-1\\1\end{pmatrix}$ is not found by the `eig` command, but can obtained by properly combining the last two eigenvectors of the two matrices - for multiple eigenvectors only *one* basis of the eigenspace can be computed, and this basis is not unique.

**Remark 3.1.34.** The problem of not finding proper joint eigenvectors has been recognized and solved in [Möller and Tenberg, 2001]; here we give a more symmetric modification of their algorithm from [Sauer, 2018a]. Note, however, that the algorithm by Möller and Tenberg can even be extended to zero finding for non-radical ideals.

The idea of the algorithm that computes and relates the common eigenvectors of the commuting family of companion matrices for a *radical* zero dimensional ideal $\mathscr{I}$, we make use of the convenient approach to identify a matrix $A \in \mathbb{C}^{n\times m}$ with the at most $m$-dimensional subspace $A\mathbb{C}^m$ of $\mathbb{C}^n$ spanned by the columns of the matrix. Since an element in intersection of the subspaces $A \cap B$ must be of the form

$$Av = Bw \qquad \Leftrightarrow \qquad Av - Bw = 0 \qquad \Leftrightarrow \qquad \big(A|-B\big)\begin{pmatrix}v\\w\end{pmatrix}$$

the intersection can be determined as the kernel of the COMPOUND MATRIX $\big(A|-B\big)$, where good old SVD comes in handy again and immediately gives Algorithm 3.1.3. Some more information about adapting the threshold in this method can be found in [Sauer, 2018a].
Next, a little bit of convenient notation.

**Definition 3.1.35.** For a diagonalizable matrix $A \in \mathbb{C}^{n\times z}$ we denote by $\Lambda(A) = (\lambda, E)$, $\lambda \in \mathbb{C}^n$, $E = [e_1, \ldots, e_n]$ the EIGENSTRUCTURE of $A$, that is,

$$Ae_j = \lambda_j e_j, \qquad j = 1, \ldots, n.$$

**Algorithm 3.1.3** SUBSPACE INTERSECTION: $A \in \mathbb{C}^{n\times a}$, $B \in \mathbb{C}^{n\times b}$, threshold $\tau > 0$.

1: Compute SVD $U\Sigma V^* \leftarrow \big(A \mid -B\big)$
2: Compute numerical rank $r \leftarrow \max\{k : \sigma_k > \tau\}$
3: Set

$$C \leftarrow \frac{1}{2}\big(AV_{1:a,r+1:a+b} + BV_{a+1:a+b,r+1:a+b}\big) \in \mathbb{C}^{n\times a+b-r}$$

4: **Result:** $C = A \cap B$.

We write the eigenvalues with multiplicities as $(\hat{\lambda}, \mu) \in \mathbb{C}^m \times \mathbb{N}^m$ for some $m \le n$ which means that $\lambda$ is a permutation of

$$(\underbrace{\hat{\lambda}_1, \ldots, \hat{\lambda}_1}_{\mu_1}, \ldots, \underbrace{\hat{\lambda}_m, \ldots, \hat{\lambda}_m}_{\mu_m})$$

and $\hat{\lambda}_j \neq \hat{\lambda}_k$, $j \neq k$.

**Remark 3.1.36.** The eigenstructure $\Lambda(A)$ is what `matlab` or `octave` computes with the `eig` command.

The method to compute the joint eigenspace and thus the zeros is given in Algorithm 3.1.4.

**Algorithm 3.1.4** EIGENSPACE INTERSECTION: $M_j\mathbb{C}^{n\times n}$ jointly diagonalizable.

1: $\big((\hat{\lambda}, \mu), E\big) := (\lambda, E) \leftarrow \Lambda(M_1)$
2: Partition

$$V = \big(V_1 \mid \ldots \mid V_m\big) \leftarrow EP, \qquad P \text{ permutation},$$

such that

$$M_j V_k = \hat{\lambda}_k V_k, \qquad k = 1, \ldots, m$$

3: $\ell \leftarrow m$
4: **for** $j = 2, \ldots, s$ **do**
5: $(\lambda, E) \leftarrow \Lambda\big(V^{-1} M_j V\big)$
6: Partition as $E = \big(E_1 \mid \ldots \mid E_m\big)$
7: $V_k \leftarrow \big(V_k \cap VE_1 \mid \ldots \mid V_k \cap VE_m\big)$, $k = 1, \ldots, \ell$
8:

$$\ell \leftarrow \sum_{k=1}^{\ell} \#\{t : \dim(V_k \cap VE_t) \ge 1\} \tag{3.1.22}$$

9: **end for**
10: **Result:** $C = A \cap B$.

**Remark 3.1.37** (Eigenspace intersection method)**.**

1. The partitioning requested in Step 2 of Algorithm 3.1.4 comes for free when using the function `eig` in `matlab` or `octave` as it orders the eigenvalues according to some criterion.

2. The variable $\ell$ encodes the number of blocks in the actual partition.

3. The PREDIAGONALIZATION by means of $V$ in Step 5 is not necessary, but improves the computational performance: whenever a column of $V$ is already an eigenvalue of $M_j$ for an eigenvalue $\lambda$, we get that

$$V^{-1}M_jV = \begin{pmatrix} * & \dots & * & 0 & * & \dots & * \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ * & \dots & * & 0 & * & \dots & * \\ 0 & \dots & 0 & \lambda & 0 & \dots & 0 \\ * & \dots & * & 0 & * & \dots & * \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ * & \dots & * & 0 & * & \dots & * \end{pmatrix},$$

which significantly eases the work of the `eig` function.

4. The operation in (3.1.22) counts the new number of pieces in the partition of $V$. If $V_k \cap VE_j = \{0\}$ would be trivial, it can be represented by an empty matrix and thus even omitted.

**Theorem 3.1.38.** *If $M_j$, $j = 1,\dots,s$, are companion matrices of a zero dimensional radical ideal, then Algorithm 3.1.4 computes a matrix $V$ that simultaneously diagonalizes the matrices $M_j$.*

**Proof:** We first prove by induction that after $j$ steps the blocks $V_1,\dots,V_\ell$ of $V$ are exactly the nontrivial intersections of eigenspaces of $M_1,\dots,M_j$ and for $j = s$ this proves our theorem since these intersections must be one-dimensional and consist of the $\ell_z$, $z \in Z(\mathscr{I})$. For $j = 1$, the statement is exactly the setup in Step 2 of the algorithm and to advance the induction hypothesis, we note that, by definition of $E$ in step $j+1$, we have

$$M_{j+1}VE_t = V\underbrace{V^{-1}M_{j+1}VE_t}_{=\hat{\lambda}_t E_t} = \hat{\lambda}_t VE_t, \qquad t = 1,\dots,m,$$

hence any such intersection $V_k \cap VE_t$ is a joint eigenspace of $M_{j+1}$ and, by induction, also one of $M_1,\dots,M_j$, which completes the induction.

Moreover, $E$ and $V$ are nonsingular, hence $\mathbb{C}^n = VE_1 \oplus \cdots \oplus VE_m$ and therefore

$$V_k = V_k \cap \mathbb{C}^n = V_k \cap \left(\bigoplus_{t=1}^{m} VE_t\right) = \bigoplus_{t=1}^{m}(V_k \cap VE_t),$$

which yields $\mathbb{C}^n = V_1 \oplus \cdots \oplus V_\ell$ at each step $j$ and proves that $V$ contains *all* intersections. □

**Example 3.1.39.** We continue Example 3.1.33 and start with the decomposition

$$M_x \sim \left(\begin{pmatrix}1\\0\\0\end{pmatrix}, \begin{pmatrix}0 & \sqrt{2} & 0\\ 1 & -\sqrt{2} & 0\\ 0 & 0 & 1\end{pmatrix}\right) \quad \Rightarrow \quad V = \left[\begin{array}{c|cc} 0 & \sqrt{2} & 0\\ 1 & -\sqrt{2} & 0\\ 0 & 0 & 1\end{array}\right].$$

With the direct decomposition[5]

$$M_y \sim \left(\begin{pmatrix}1\\0\\0\end{pmatrix}, \begin{pmatrix}0 & \sqrt{2} & 0\\ 0 & 0 & 1\\ 1 & -\sqrt{2} & 0\end{pmatrix}\right) \quad \Rightarrow \quad [VE_1 \,|\, VE_2] = \left[\begin{array}{c|cc} 0 & \sqrt{2} & 0\\ 0 & 0 & 1\\ 1 & -\sqrt{2} & 0\end{array}\right]$$

[5] Without the prediagonalization step.

We compute the intersections

$$V_1 \cap VE_1 = \{0\}, \qquad V_1 \cap VE_2 = \operatorname{span}\begin{pmatrix}0\\1\\0\end{pmatrix}, \qquad V_2 \cap VE_1 = \operatorname{span}\begin{pmatrix}0\\0\\1\end{pmatrix}, \qquad V_2 \cap VE_1 = \operatorname{span}\begin{pmatrix}\sqrt{2}\\-\sqrt{2}\\-\sqrt{2}\end{pmatrix},$$

and found the three normalized eigenvectors $\begin{pmatrix}0\\1\\0\end{pmatrix}, \begin{pmatrix}0\\0\\1\end{pmatrix}, \frac{1}{\sqrt{3}}\begin{pmatrix}1\\-1\\-1\end{pmatrix}$ and

$$V = \begin{pmatrix}0 & 0 & \sqrt{3}\\ 1 & 0 & -\sqrt{3}\\ 0 & 1 & -\sqrt{3}\end{pmatrix}$$

has the property that

$$V^{-1} M_x V = \begin{pmatrix}1 & & \\ & 0 & \\ & & 0\end{pmatrix}, \qquad V^{-1} M_x V = \begin{pmatrix}0 & & \\ & 1 & \\ & & 0\end{pmatrix},$$

from which we can read off the zeros $(1,0)$, $(0,1)$ and $(0,0)$.

## 3.2 Zeros and their multiplicity

We now come to one of the fundamental chapters in this lecture, namely a precise description of multiple zeros of polynomials. To my knowledge, this is due to Gröbner [Gröbner, 1939], see also [Gröbner, 1970].

### 3.2.1 Invariances and dualities

We begin with some fundamental concepts.

**Definition 3.2.1** (Invariances)**.** A set $\mathscr{P} \subset \Pi = \mathbb{K}[x]$ is called

1. SHIFT INVARIANT if
$$f \in \mathscr{P} \qquad \Rightarrow \qquad f(\cdot + y) \in \mathscr{P}, \quad y \in \mathbb{K}^s, \tag{3.2.1}$$
2. $D$-INVARIANT or DIFFERENTIATION INVARIANT if
$$f \in \mathscr{P} \qquad \Rightarrow \qquad q(D) f \in \mathscr{P}, \quad q \in \Pi. \tag{3.2.2}$$
3. By $\tau_y$, $y \in \mathbb{K}^s$, we denote the SHIFT OPERATOR, $\tau_y f := f(\cdot + y)$.

**Remark 3.2.2.** Specifying $y = 0$ or $q = 1$, we immediately see that the direction "$\Leftarrow$" is also valid in (3.2.1) and (3.2.2), respectively.

It is worthwhile to note that the two invariance concepts are in fact the same one.

**Proposition 3.2.3.** *A subspace $\mathscr{P} \subset \Pi$ is shift invariant if and only it is $D$-invariant.*

**Proof:** If $\mathscr{P}$ is shift invariant, then all the polynomials $\frac{1}{h}(\tau_{hy}-I)f$, $h\in\mathbb{R}_+$, belong to the finite dimensional closed space $\mathscr{P}\cap\Pi_{\deg f-1}$ by means of (3.2.3). Therefore, the limit $D_y f$, the DIRECTIONAL DERIVATIVE with respect to $y$ of $f$, also belongs to $\mathscr{P}$. Conversely, the TAYLOR FORMULA

$$\tau_y f = f(\cdot+y) = \sum_{\alpha\in\mathbb{N}_0^s} \frac{\partial^{|\alpha|} f}{\partial x^\alpha}(\cdot)\,\frac{y^\alpha}{\alpha!}$$

immediately implies that $\tau_y f$ is a finite linear combination of derivatives of $f$, hence shift invariance follows from $D$-invariance. □

**Exercise 3.2.1** Prove the formula

$$f(x+y) = \sum_{|\alpha|\le\deg f} f_\alpha \sum_{\beta\le\alpha} \binom{\alpha}{\beta} x^\beta y^{\alpha-\beta}, \qquad x,y\in\mathbb{K}^s. \tag{3.2.3}$$

♢

**Lemma 3.2.4.** *If $\mathscr{P}\subset\Pi$ is a nontrivial $D$-invariant space, then $1\in\mathscr{P}$.*

**Proof:** Choose $0\neq f\in\mathscr{P}$, then $f(D)f$ is a nonzero constant that belongs to $\mathscr{P}$ since the space is $D$-invariant. □

Next, we extend "our" inner product.

**Definition 3.2.5.** By $(\cdot,\cdot)_z$, $z\in\mathbb{K}^s$, we denote the bilinear form

$$(f,g)_z := \big(f(D)g\big)(z) = \big(f,\tau_z g\big). \tag{3.2.4}$$

Since

$$(f,g)_z = (f,\tau_z g) = (\tau_z g, f) = \big((\tau_z g)(D)f\big)(0) = \big((\tau_z g)(D)\tau_{-z}f\big)(z) = \big(\tau_z g,\tau_{-z}f\big)_z,$$

the bilinear form is *not* symmetric for $z\neq 0$. But still it has the fundamental property that

$$(fq,g)_z = (f(D)q(D)g)(z) = \big(f,q(D)g\big)_z. \tag{3.2.5}$$

**Proposition 3.2.6.** *For $z\in\mathbb{K}^s$ let $\mathscr{F},\mathscr{G}\subseteq\Pi$ be related by*

$$\mathscr{F} := \ker(\cdot,\mathscr{G})_z := \big\{f\in\Pi : (f,\mathscr{G})=0\big\}, \qquad \mathscr{G} := \ker(\mathscr{F},\cdot)_z := \big\{g\in\Pi : (\mathscr{F},g)=0\big\}. \tag{3.2.6}$$

*Then $\mathscr{F}$ is an ideal if and only if $\mathscr{G}$ is $D$-invariant.*

**Proof:** $\mathscr{F}$ and $\mathscr{G}$ are both closed under addition since the form is bilinear. Let $\mathscr{F}$ be an ideal and $q\in\Pi$. Then, for any $g\in\mathscr{G}$

$$0 = (qf,g)_z = (f,q(D)g)_z, \qquad f\in\mathscr{F},$$

hence $\mathscr{G}$ is $D$-invariant. For the converse, we note that for $f\in\mathscr{F}$ we have that

$$0 = (f,q(D)g)_z = (qf,g)_z, \qquad g\in\mathscr{G},$$

hence $qf\in\mathscr{F}$ and thus $\mathscr{F}$ is an ideal. □

**Definition 3.2.7.** A pair $\mathscr{F},\mathscr{G}\subseteq\Pi$ for which (3.2.6) holds is called a DUAL PAIR of subspaces.

### 3.2.2 Multiple zeros

First a small bit of terminology.

**Definition 3.2.8.** The POINT EVALUATION FUNCTIONAL $\delta_z : \Pi \to \mathbb{K}$, $z \in \mathbb{K}^s$ is defined as $\delta_z f = f(z)$.

Now we are in position to characterize zero dimensional ideals in terms of zeros and their multiplicities.

**Theorem 3.2.9.** *Let $\mathbb{K}$ be an algebraically closed field*[6]*. An ideal $\mathscr{I} \subseteq \Pi$ is zero dimensional if and only if there exist a finite set $Z(\mathscr{I}) \subset \mathbb{K}^s$ and finite dimensional $D$-invariant spaces $\mathscr{Q}_z$, $z \in Z(\mathscr{I})$, such that*

$$\mathscr{I} = \bigcap_{z \in Z(\mathscr{I})} \ker \delta_z \circ \mathscr{Q}_z(D) = \left\{ f \in \Pi : q(D) f(z) = 0,\, q \in \mathscr{Q}_z, z \in Z(\mathscr{I}) \right\} \tag{3.2.7}$$

**Proof:** Suppose the $\mathscr{I}$ is zero dimensional. Since for any $z \in \mathbb{K}^s$ the set $\mathscr{G}^\perp(z) = \ker(\mathscr{I}, \cdot)_z$ is a $D$-invariant subset of $\Pi$. By Lemma 3.2.4 it is either trivial or contains 1 and since the latter implies that $\mathscr{I}(z) = 0$, we can conclude that

$$\mathscr{G}^\perp(z) \begin{cases} = \{0\}, & z \notin Z(\mathscr{I}), \\ \supseteq \Pi_0, & z \in Z(\mathscr{I}). \end{cases} \tag{3.2.8}$$

Set $\mathscr{Q}_z = \mathscr{G}^\perp(z)$. Then, again by Proposition 3.2.6

$$\mathscr{J} := \bigcap_{z \in Z(\mathscr{I})} \ker(\cdot, \mathscr{Q}_z)_z \tag{3.2.9}$$

is an intersection of ideals, hence an ideal, and since $(\mathscr{I}, \mathscr{Q}_z)_z = 0$, it follows that $\mathscr{I} \subseteq \mathscr{J}$ and $Z(\mathscr{I}) = Z(\mathscr{J})$. Therefore,

$$\ker(\mathscr{J}, \cdot)_z \subseteq \ker(\mathscr{I}, \cdot)_z = \mathscr{Q}_z$$

and a strict inclusion $\mathscr{I} \subset \mathscr{J}$ would imply a strict inclusion above. But this is impossible since (3.2.9) implies that $(\mathscr{J}, \mathscr{Q}_z)_z = 0$, hence $\ker(\mathscr{J}, \cdot)_z \supseteq \mathscr{Q}_z$.

For the converse, we first note that each $\mathscr{I}_z := \ker(\delta_z \circ \mathscr{Q}_z(D))$ is a primary ideal with variety $\{z\}$. Since $\mathscr{Q}_z$ is finite dimensional, it follows that

$$\deg \mathscr{Q}_z = \max_{q \in \mathscr{Q}_z} \deg q < \infty$$

and therefore $\left(q(D)\tau_{-z} f\right)(z) = 0$ for any $f \in \Pi_k^0$ for $k > \deg \mathscr{Q}_z$, so that $V_k(\mathscr{I}_z) = \Pi_k^0$ for $k > \deg \mathscr{Q}_z$ and $\Pi/\mathscr{I}_z \subseteq W(\mathscr{I}_z) \subseteq \Pi_{\deg \mathscr{Q}_z}$ is finite dimensional. This carries over to the intersections. □

**Exercise 3.2.2** Prove that the intersection of two zero dimensional ideals is zero dimensional again, in particular

$$\dim \Pi/(\mathscr{I} \cap \mathscr{J}) \leq \dim \Pi/\mathscr{I} + \dim \Pi/\mathscr{J}.$$

♢

[6] Alternatively one could request $Z \subset \overline{\mathbb{K}}$, the ALGEBRAIC CLOSURE of $\mathbb{K}$.

**Definition 3.2.10.** The space $\mathscr{Q}_z$ is called the (structural) MULTIPLICITY of the zero at $z \in Z(\mathscr{I})$. The SCALAR MULTIPLICITY is $\mu(z) := \dim \mathscr{Q}_z$. Moreover, we use the notation

$$\mathscr{I}_z := \ker \delta_z \circ \mathscr{Q}_z(D) = \left\{ f \in \Pi : \big(q(D)f\big)(z) = 0,\, q \in \mathscr{Q}_z \right\} \tag{3.2.10}$$

for the local primary components.

For more information on the dimensionality of the associated quotient spaces, it pays off to use a slight extension of Theorem 2.1.8.

**Corollary 3.2.11.** *For $f, q \in \Pi$ and $z \in \mathbb{K}^s$ we have that*

$$\big(q(D)f\big)(z) = (f, q\,e_z), \qquad e_z := \sum_{\alpha \in \mathbb{N}_0^s} \frac{z^\alpha}{\alpha!}(\cdot)^\alpha = e^{z^T(\cdot)}. \tag{3.2.11}$$

**Proof:** By (2.1.11),

$$(q(D)f)(z) = (q(D)f, e_z) = \big(f, q e_z\big).$$

□

**Proposition 3.2.12** (Dimensions)**.** *For any zero dimensional ideal $\mathscr{I}$ in a algebraically closed field $\mathbb{K}$ we have that*

$$\dim \Pi/\mathscr{I}_z = \dim \mathscr{Q}_z \tag{3.2.12}$$

*and*

$$\dim \Pi/\mathscr{I} = \sum_{z \in Z(\mathscr{I})} \dim \mathscr{Q}_z. \tag{3.2.13}$$

**Proof:** Let $P_z$ be a basis of $\Pi/\mathscr{I}_z$ and $Q_z$ a basis of $\mathscr{Q}_z$ and consider the matrix

$$G := \left( (p, q e_z) : \begin{array}{l} p \in P_z \\ q \in Q_z \end{array} \right)$$

If the rank of this matrix is smaller than $\#Q_z$ then there exists $v \in \mathbb{K}^{Q_z}$ such that $Gv = 0$, hence

$$0 = (\Pi/\mathscr{I}, (v \cdot Q_z)e_z)$$

and since any $f \in \Pi$ can be written as $\nu_{\mathscr{I}}(f) + f'$, $f' \in \mathscr{I}_z = \ker(\cdot, \mathscr{Q}_z e_z)$, it follows that

$$\big(f, (v \cdot Q_z)e_z\big) = \big(\nu_{\mathscr{I}}(f), (v \cdot Q_z)e_z\big) + \big(f', (v \cdot Q_z)e_z\big) = 0,$$

from which we conclude that $v \cdot Q_z = 0$ and thus $v = 0$. Hence, $\operatorname{rank} G = \#Q_z = \dim \mathscr{Q}_z$ which proves (3.2.12). The same argument, now taking into account the linear independence of all $qe_z$, $q \in Q_z$, $z \in Z(\mathscr{I})$, also verifies (3.2.13). □

**Example 3.2.13.** Let us consider triple zeros at $(0,0)$. One choice is that $\mathscr{Q}_0 = \Pi_1$ which leads to

$$\mathscr{I} = \langle x^2, xy, y^2 \rangle, \qquad \Pi/\mathscr{I} = \Pi_1 \tag{3.2.14}$$

or we can choose $Q_0 = \{1, p, q\}$ where $p$ is an affine polynomial of the form[7] $p = ax + by$ and $q = ux^2 + vxy + wy^2 + cx + dy$. To ensure $D$-invariance, we have to consider

$$\frac{\partial q}{\partial x} = 2ux + vy + c, \qquad \frac{\partial q}{\partial y} = vx + 2wy + d$$

[7] We can drop the constant part as it is covered by the constant in the basis.

and choose the parameters in such a way that both derivatives are contained in span $\{1, p\}$, yielding[8]

$$\frac{2u}{a} = \frac{v}{b}, \qquad \frac{v}{a} = \frac{2w}{b},$$

hence, if we write $v = \alpha ab$, we get $u = \frac{\alpha}{2}a^2$ and $w = \frac{\alpha}{2}b^2$, yielding

$$q(x,y) = \frac{\alpha}{2}\left(a^2x^2 + 2abxy + b^2y^2\right) + cx + dy = \frac{\alpha}{2}\left(ax + by\right)^2 + cx + dy = \frac{\alpha}{2}p(x,y)^2 + cx + dy,$$

so the general case is

$$\mathscr{Q}_0 = \operatorname{span}\{1, p, p^2 + \ell\}, \qquad \ell \in \Pi_1^0.$$

If $q \in \Pi_1^0$ is such that $(q,p) = 0$, i.e., $q(x,y) = bx - ay$, then

$$\mathscr{I} = \langle q, p^3 \rangle, \qquad \Pi/\mathscr{I} = \operatorname{span}\{1, p, p^2\},$$

with the special cases

$$\mathscr{I} = \langle x^3, y \rangle, \quad \Pi/\mathscr{I} = \operatorname{span}\{1, x, x^2\} \qquad \text{and} \qquad \mathscr{I} = \langle x, y^3 \rangle, \quad \Pi/\mathscr{I} = \operatorname{span}\{1, y, y^2\}$$

of pure partial derivatives.

[8] The special cases that either $a = 0$ and $b = 0$ have to be considered separately, but they are simpler as they lead to $\mathscr{I} = \langle x, y^3 \rangle$ and $\mathscr{I} = \langle x^3, y \rangle$, respectively.

# Interpolation | 4

*Il est manifeste que l'interpolation des fonctions de plusiers variables ne demande aucun principe nouveau, car dans tout ce qui précède le fait que la variable indépendante était unique n'a souvent joué aucun rôle.*

*It is clear that the interpolation of functions of several variables does not demand any new principles because in the above exposition the fact that the variable was unique has not played frequently any role.*

(H. Andoyer in [Andoyer, 1906])

---

INTERPOLATION can be seen as a RECOVERY PROBLEM:

Given SITES $\mathscr{X} \subset \mathbb{K}^s$, $\#\mathscr{X} < \infty$, and $y \in \mathbb{K}^{\mathscr{X}}$ find $f \in \Pi$ such that

$$f(\mathscr{X}) = y, \qquad \text{i.e.,} \qquad f(x) = y_x, \quad x \in \mathscr{X}. \tag{4.0.1}$$

The task in (4.0.1) is to reconstruct or *recover* a function $f$ in a structured way from finite information on the function, in this case from the value at certain points.

Multivariate interpolation is a fairly recent topic, especially compared to the fact that the univariate case is already covered in [Newton, 1687] and that, according to [Bauschinger, 1900] the name has been coined by J. Wallis as early as even 1655.

The oldest work on multivariate interpolation can be found in [Jacobi, 1835, Kronecker, 1866], for details see [Gasca and Sauer, 2000b]. Systematic studies began around 1900 in the context of algebraic geometry, numerically it was considered in [Radon, 1948]. In all what follows, we only consider the case of *finitely many* interpolation conditions, avoiding concepts like TRANSFINITE INTERPOLATION or interpolation along curves.

## 4.1 Basic aspects

We begin by collecting some very elementary facts about polynomial interpolation and mainly giving a series of definitions and very straightforward results.

### 4.1.1 Terminology

The simplest type of interpolation problem is the one defined in (4.0.1) that depends only on point evaluations.

**Definition 4.1.1.** An interpolation problem is called a LAGRANGE INTERPOLATION problem if the result depends on *function values* only, i.e., if it is of the form (4.0.1).

**Lemma 4.1.2.** *The Lagrange interpolation problem is always solvable.*

**Proof:** Use

$$f = \sum_{x \in \mathscr{X}} y_x \prod_{x' \neq x} \frac{(x - x')^H (\cdot - x')}{\|x - x'\|_2^2}. \tag{4.1.1}$$

□

In general, we can start with a *finite* set $\theta : \Pi \to \mathbb{K}$ of linear functionals defined on the polynomials; linear independence means that

$$0 = \sum_{\theta \in \Theta} c_\theta \theta(f) = 0, \quad f \in \Pi \qquad \Leftrightarrow \qquad c_\theta = 0, \quad \theta \in \Theta.$$

**Definition 4.1.3.** Let $\Theta \subset \Pi'$, $\#\Theta < \infty$. The GENERALIZED INTERPOLATION PROBLEM consist of finding, for any $y \in \mathbb{K}^\Theta$ a polynomial $f \in \Pi$ such that

$$\Theta(f) = y, \qquad \text{i.e.,} \qquad \theta(f) = y_\theta, \quad \theta \in \Theta. \tag{4.1.2}$$

The functionals in $\Pi'$ can be easily embedded into the formal power series since

$$\theta(f) = (f, e_\theta), \qquad e_\theta(x) := \sum_{\alpha \in \mathbb{N}_0^s} \frac{\theta\left((\cdot)^\alpha\right)}{\alpha!} x^\alpha, \tag{4.1.3}$$

see the proof of Theorem 2.1.8.

Now we can classify generalized schemes in various ways, originally due to Birkhoff [Birkhoff, 1979].

**Definition 4.1.4** (Ideal interpolation)**.** A generalized interpolation problem is called an IDEAL INTERPOLATION problem if its kernel

$$\ker \Theta := \{f \in \Pi : \Theta(f) = 0\}$$

is an ideal.

**Remark 4.1.5.** Lagrange interpolation is ideal interpolation since $\{f \in \Pi : f(\mathscr{X}) = 0\}$ is an ideal.

By Theorem 3.2.9 any zero dimensional ideal $Z(\mathscr{I})$ can be written as the kernel of $D$-invariant spaces of differential conditions evaluated at $Z(\mathscr{I})$. Hence any ideal interpolation problem can be interpreted as a Hermite interpolation problem where *consecutive* derivatives have to be interpolated.

**Definition 4.1.6.** A HERMITE INTERPOLATION problem consists of the conditions

$$(q(D)f)(x) = y_{q,x}, \qquad q \in Q_x,\, x \in \mathscr{X}, \tag{4.1.4}$$

where $Q_x$ is a basis of the $D$-invariant subspace $\mathscr{Q}_x \subset \Pi$. A generalized interpolation problem that is not ideal is called a HERMITE-BIRKHOFF INTERPOLATION problem.

**Remark 4.1.7.** $D$-invariance takes care of the derivatives being consecutive.

**Example 4.1.8.** A simple Hermite-Birkhoff interpolation problem is to interpolate $\partial f / \partial x$ at some point, but not the function value.

**Definition 4.1.9.** A linear subspace $\mathscr{P} \subseteq \Pi$ is called an INTERPOLATION SPACE for the generalized interpolation problem with respect to $\Theta \subset \Pi'$ if for any $y \in \mathbb{K}^\Theta$ there exists $p \in \mathscr{P}$ such that $\Theta(p) = y$. It is called a UNIQUE INTERPOLATION SPACE if the INTERPOLANT $p$ is unique.

### 4.1.2 Linear algebra and the difference to the univariate case

As long as linear functionals are involved, we can use obvious linear algebra to describe the solvability of interpolation problems. The following concept is fundamental for that purpose.

**Definition 4.1.10.** For $P \subseteq \Pi$ and $\Theta \subset \Pi'$, the associated VANDERMONDE MATRIX is defined as

$$V(P,\Theta) := \begin{pmatrix} \theta(p) : & \theta \in \Theta \\ & p \in P \end{pmatrix} \tag{4.1.5}$$

and describes interpolation with respect to $\Theta$ on $\mathscr{P}$. In the case of Lagrange interpolation problems we simply write $V(P,\mathscr{X})$ and in the case of Hermite interpolation we use $V(P,(\mathscr{X},Q))$ where $Q = (Q_x : x \in \mathscr{X})$ is a vector of bases for the $D$-invariant spaces $\mathscr{Q}_x$.

With the Vandermonde matrix we can write the interpolation problem as

$$y = V(P,\Theta)v, \qquad v \in \mathbb{K}^P, \tag{4.1.6}$$

as then $\Theta(v \cdot P) = y$. Standard linear algebra gives us the following observation.

**Theorem 4.1.11.** *Let $P$ be a basis for $\mathscr{P} \subseteq \Pi$ and $\Theta \subset \Pi'$ be a finite set of linearly independent functionals. Then*

1. *$\mathscr{P}$ is an interpolation space for $\Theta$ if and only if* $\operatorname{rank} V(P,\Theta) \geq \#\Theta$,
2. *$\mathscr{P}$ is an unique interpolation space for $\Theta$ if and only if $V(P,\Theta)$ is invertible.*

Even if this theorem is almost trivial as it is only a reformulation of the problem, it gives us some insight into the differences between the univariate and the multivariate case. In the univariate case, as shown in Theorem 1.2.1, any $n+1$ pairwise distinct points could be uniquely interpolated by $\Pi_n$. This makes $\Pi_n$ a so-called HAAR SPACE where interpolation is a matter of counting and dimension only. This is lost in higher dimensions and it is not even a matter of polynomials.

**Example 4.1.12.** If $X \subset \mathbb{R}^2$ contains an open set or a branch, then there exists *no* Haar space for functions defined on $X$.

**Proof:** Given $\mathscr{X}$ with $\#\mathscr{X} \geq 2$, we choose $x, x' \in \mathscr{X}$, $x \neq x'$, and two continuous functions $u, v : [0,1] \to X$ such that

$$u(0) = v(1) = x, \qquad u(1) = v(0) = x', \qquad u(t) \neq v(t), \quad t \in [0,1],$$

and $u(t), v(t) \notin \mathscr{X} \setminus \{x, x'\}$. In other words: $u$ and $v$ switch $x$ and $x'$ continuously without violating the fact that the points are different. Now let $\Phi$ with $\#\Phi = \#\mathscr{X}$ an arbitrary linearly independent set of functions $X \to \mathbb{R}$.

$$D(t) = \det\begin{pmatrix} \phi(y) : & \phi \in \Phi \\ & y \in \{u(t), v(t)\} \cup \big(\mathscr{X} \setminus \{x, x'\}\big) \end{pmatrix}, \qquad t \in [0,1],$$

is continuous in $t$ and satisfies $D(0) = -D(1)$, due to which there must exist $t^* \in [0,1]$ such that $D(t^*) = 0$, so that interpolation at $\{u(t^*), v(t^*)\} \cup \big(\mathscr{X} \setminus \{x, x'\}\big), v(t^*), x_3, \ldots, x_n$ is impossible. This procedure is shown graphically in Fig. 4.1.1 □

In fact, it turns out that Haar spaces exist in the univariate case only:

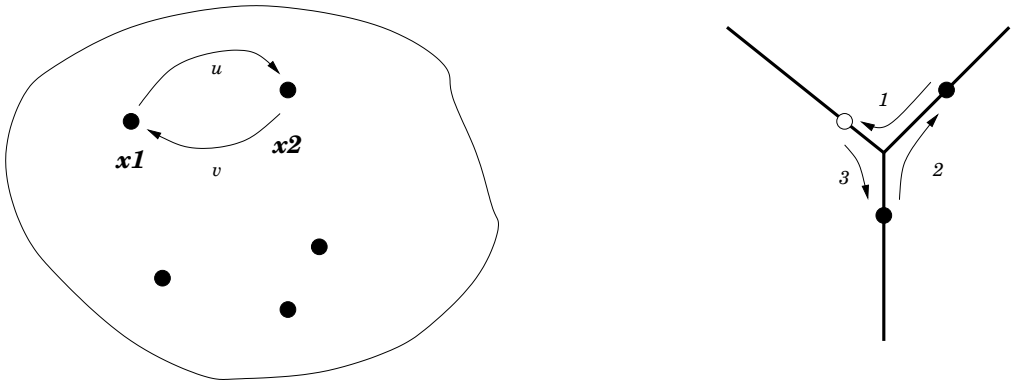



Figure 4.1.1: Switching the points $x = x_1$ and $x' = x_2$, somewhere in between interpolation has to fail. On the right hand side the procedure for a branch, cf. [Lorentz, 1966, S. 25]), which works like shuffling trains on a railroad track.

**Theorem 4.1.13** (MAIRHUBER'S THEOREM, [Mairhuber, 1956])**.** *For a compact metric space $X$ there exist nontrivial[1] Haar spaces if and only if $X$ is homeomorphic to a compact subset of the* TORUS $\mathbb{T}$.

Example 4.1.12 has some important consequences that we list next.

**Remark 4.1.14** (Multivariate interpolation)**.**

1. For each subspace $\mathscr{P} \subset \Pi$ of polynomials with basis $P$ there exist configurations $\mathscr{X}$ of sites such that $\det V(P, \mathscr{X}) = 0$, hence interpolation is not solvable in general on this set. Interpolation is not a matter of *counting* but of *geometry*.

2. Each nontrivial subspace $\mathscr{P}$ with basis $P$ has a dual set of points $\mathscr{X}$, $\#\mathscr{X} = \#P$, for which $\det V(P, \mathscr{X}) \neq 0$. This is easily proved by induction on $\#P$, where $\#P = 1$ is simply the fact that a basis element is by definition not identically zero. For the induction step, we pick some $p \in P$, find a point $x$ where $p(x) \neq 0$ and apply the induction hypothesis to
$$P' = \left\{ p' - \frac{p'(x)}{p(x)} p : p' \in P \setminus \{p\} \right\}$$
to obtain a set[2] $\mathscr{X}'$ such that $\det V(P', \mathscr{X}') \neq 0$. Then, with $\mathscr{X} = \{x\} \cup \mathscr{X}'$,
$$\det V(P, \mathscr{X}) = \pm p(x) \det \big( V(P', \mathscr{X}') A \big) \neq 0,$$
where $A$ is the nonsingular matrix that switches between the bases $P$ and $\{p\} \cup P'$.

3. If $\#P = \#\mathscr{X}$ such that $\det V(P, \mathscr{X}) = 0$, we know that $\det V(P, \cdot)$ is a nonzero polynomial in the components of the sites and thus only vanishes on a set of measure zero. In other words: polynomials are an *almost Haar space*.

4. Nevertheless it is the set of measure zero that causes trouble and can make polynomial interpolation problems arbitrarily ill-conditioned even if the points are well-separated.

[1] $\Phi = \{0\}$ and $\Phi = \mathbb{C}$ are Haar spaces but not interesting.
[2] Since $P'(x) = 0$ we cannot have $x \in \mathscr{X}'$.

In the univariate case any Hermite interpolation problem can be expressed as the limit of Lagrange interpolation problems with COALESCING POINTS which means that in the limit some of the interpolation points collapse into a single one and the direction of collapsing[3] determines the directional derivative. This does no more in such a straightforward way in the multivariate case.

**Example 4.1.15.** For any $h > 0$, the Lagrange interpolation problem at

$$\mathscr{X}_h := \{(0,0),(h,0),(0,h),(1,1),(1+h,1),(1,1+h)\}$$

has a unique solution in $\Pi_2$, as can be easily seen from considering the Vandermonde matrix

$$V(P,\mathscr{X}_h) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & h & 0 & h^2 & 0 & 0 \\ 1 & 0 & h & 0 & 0 & h^2 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1+h & 1 & (1+h)^2 & 1+h & 1 \\ 1 & 1 & 1+h & 1 & 1+h & (1+h)^2 \end{pmatrix}, \qquad P = \{1, x, y, x^2, xy, y^2\}.$$

whose determinant is $-4h^5$. The limit problem, on the other hand, would be a Hermite interpolation with

$$\mathscr{X} = \{(0,0),(1,1)\}, \qquad \mathscr{Q}_{(0,0)} = \mathscr{Q}_{(1,1)} = \Pi_1$$

which is not solvable in $\Pi_2$ as it defines 4 conditions on the line connecting the two interpolation points.

## 4.2 Interpolation constructions

In view of the examples of the preceding section it becomes clear that in several variables even the question when an interpolation problem is (uniquely) solvable depends on the space $\mathscr{P}$ and the set of functionals $\Theta$ or, in the simpler case of Lagrange interpolation problems, the set $\mathscr{X}$ of sites. There are essentially two directions of research, namely

1. given a space $\mathscr{P}$ determine $\mathscr{X}$ or even $\Theta$ such that the associated interpolation problem is (uniquely) solvable,
2. given a set $\Theta$ of functionals or $\mathscr{X}$ of sites, find a space $\mathscr{P}$ that allows for (unique) interpolation.

There are even examples of constructions that do both at the same time, cf. [Gasca and Sauer, 2000a], but mostly the above two approaches persist as can be seen in surveys about multivariate polynomial interpolation like [Gasca and Sauer, 2000c, Gasca and Sauer, 2000b, Lorentz, 2000, Sauer, 2006].

### 4.2.1 Constructing point sets

Since historically people started by constructing interpolation sets, we will follow the chronology here and start with two classical constructions due to Radon[4] [Radon, 1948] and Chung and Yao and comment on some of their connections. Both constructions are based on hyperplanes and both give sites $\mathscr{X}$ that allow for unique interpolation in the total degree space $\Pi_n$ for some $n \geq 0$.

[3] Yes, this is more complex in the multivariate situation.

[4] Yes, the guy with the famous transform in computerized tomography.

**Definition 4.2.1.** A HYPERPLANE $H \subset \mathbb{K}$ is the zero set of an affine function $h(x) = v^T x + c$, that is

$$H = \left\{ x \in \mathbb{K}^s : h(x) = 0 \right\}. \tag{4.2.1}$$

We make the convention that the NORMAL VECTOR $v$ of the hyperplane is NORMALIZED such that $\|v\|_2 = 1$. This defines it up to sign.

**Remark 4.2.2.** Since we aim for *unique* interpolation in $\Pi_n$, the associated set $\mathcal{X}_n$ of sites must satisfy $\#\mathcal{X}_n = \dim \Pi_n = \binom{n+s}{s}$.

Radon's construction is inductive on $n$ and $s$, taking into account that the case $n = 0$ is really simple: nonzero constant function interpolate nicely at a single point. Univariate interpolation on which the recurrence by $s$ is based is also simple as the points only have to be distinct.

Now suppose that a set $\mathcal{X}_n$ for $\Pi_n$ has been constructed, then one chooses a hyperplane $H$ such that $H \cap \mathcal{X}_n = \emptyset$ and $r_d = \binom{n+s}{s-1}$ points $\mathcal{X}^0_{n+1}$ on this hyperplane that allow for interpolation by $\Pi_{n+1}$ in $s-1$ variables. Setting $\mathcal{X}_{n+1} := \mathcal{X}_n \cup \mathcal{X}^0_{n+1}$, we obtain the set we are looking for.

**Theorem 4.2.3.** *The set $\mathcal{X}_{n+1}$ constructed above admits unique interpolation for $\Pi_{n+1}$.*

**Proof:** We construct the LAGRANGE FUNDAMENTAL POLYNOMIALS $\ell_x \in \Pi_{n+1}$ with $\ell_x(x') = \delta_{x,x'}$, $x, x' \in \mathcal{X}_{n+1}$, which also proves the existence of a unique interpolant for any values to be interpolated.

For $x \in \mathcal{X}_n$ we use the fact that $h(x) \neq 0$ by definition of $H$ and the fundamental polynomials $\ell'_x \in \Pi_n$ with respect to $\mathcal{X}_N$ to define

$$\ell_x := \frac{h}{h(x)} \ell'_x, \qquad x \in \mathcal{X}_n$$

with the requested properties.

For $x \in \mathcal{X}^0_{n+1}$ we write the points as $\mathcal{X}^0_{n+1} := y^* + V\mathcal{Y}_{n+1}$, where the columns of $V \in \mathbb{K}^{s \times s-1}$ complete $v$ to an orthonormal basis of $\mathbb{K}^s$. The polynomials

$$\ell'_x := \ell^0_x \left( V^H (\cdot - y^*) \right), \qquad \ell^0_x \in \mathbb{K}[y_1, \dots, y_{s-1}], \qquad x \in \mathcal{X}^0_{n+1},$$

then satisfy $\ell'_x(x') = \delta_{x,x'}$, $x, x' \in \mathcal{X}^0_{n+1}$ and can be easily transformed into the Lagrange fundamental polynomials

$$\ell_x := \ell'_x - \sum_{x' \in \mathcal{X}_n} \ell'_x(x')\, \ell_{x'} \in \Pi_{n+1}, \qquad x \in \mathcal{X}^0_{n+1},$$

which completes the fundamental basis and the proof. □

**Remark 4.2.4.** The original construction given by Radon in [Radon, 1948] only considered the case $s = 2$ and distinct points on added lines.

The other construction by Chung and Yao [Chung and Yao, 1977] is, in some way, an almost straightforward formulation of simple factorizable fundamental polynomials.

**Definition 4.2.5.** A set $\mathcal{X} \subset \mathbb{K}^s$ is said to satisfy the GEOMETRIC CHARACTERIZATION of degree $n$ if for any $x \in \mathcal{X}$ there exist hyperplanes $H_{x,j}$, $j = 1, \dots, n$, such that

$$x \notin \bigcup_{j=1}^{n} H_{x,j}, \qquad \mathcal{X} \setminus \{x\} \subset \bigcup_{j=1}^{n} H_{x,j}. \tag{4.2.2}$$

$\mathcal{X}$ is called a GC SET of degree $n$ if it satisfies the geometric characterization of degree $n$.

**Theorem 4.2.6.** *Any set $\mathcal{X}$ that satisfies the geometric characterization of degree $n$ admits unique interpolation in $\Pi_n$.*

**Proof:** By

$$\ell_x = \prod_{j=1}^{n} \frac{h_{x,j}}{h_{x,j}(x)}, \qquad x \in \mathcal{X},$$

the Lagrange fundamental polynomials are given explicitly. □

**Definition 4.2.7.** A collection $\mathcal{H}_n := \{H_1, \ldots, H_n\}$ of hyperplanes is said to be IN GENERAL POSITION if any $s$ of them intersect in a single point:

$$\bigcap_{H \in \mathcal{H}} H = \{x_{\mathcal{H}}\}, \qquad \mathcal{H} \in \binom{\mathcal{H}_n}{s} := \left\{\mathcal{H} \subseteq \mathcal{H}_n : \#\mathcal{H} = s\right\}.$$

The intersection points points $x_{\mathcal{H}}$, $\mathcal{H} \in \binom{\mathcal{H}_{n+s}}{s}$ are called a NATURAL LATTICE of degree $n$.

**Proposition 4.2.8.** *Any natural lattice of degree $n$ is a GC set.*

**Proof:** Fixing $\mathcal{H} \in \binom{\mathcal{H}_{n+s}}{s}$, we only have to note that for any $x_{\mathcal{H}'} \in \mathcal{X} \setminus \{x_{\mathcal{H}}\}$ there exists a hyperplane $H' \in \mathcal{H}_{n+s} \setminus \mathcal{H}$ with $x_{\mathcal{H}'} \in H'$, which is exactly the geometric characterization. □

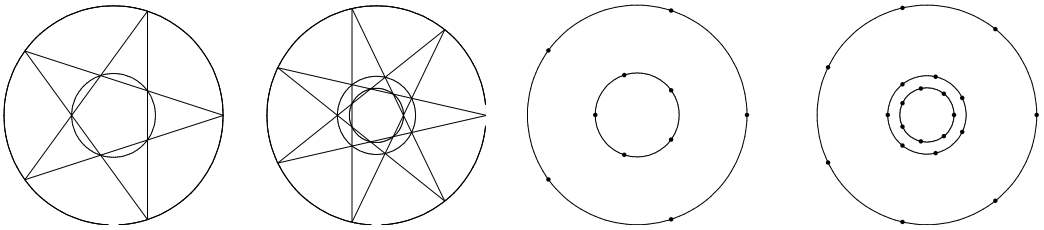


Figure 4.2.2: Natural lattices with star-shaped intersections, showing the hyperplanes *(left)* and the points *(right)* for degree 4 and degree 6.

If one tries to construct natural lattices, it turns out that they easily become quite large and "flat". It is, however, possible to construct such lattices in two variables by means of regular polyhedra, yielding point configurations that can be given explicitly and are located on concentric circles, cf. [Sauer and Xu, 1996]. This triggered quite a few constructions, most remarkably the so-called PADUA POINTS[5] [Bos et al., 2007, Caliari et al., 2006] in 2D that provide very good numerical conditioning in the sense that their Lebesgue constant grows very moderately.

In addition, there is a famous conjecture in bivariate polynomial interpolation due to Gasca and Maeztu [Gasca and Maeztu, 1982].

**Conjecture 4.2.9.** *In 2D any GC set can be obtained by the Radon construction, that is, if $\mathcal{X}$ is a GC set of degree $n$ then there exist lines containing $n, n-1, \ldots, 1$ points.*

The case $n = 1$ of Conjecture 4.2.9 is trivial, the case $n = 2$ is a simple counting argument given in [Gasca and Maeztu, 1982]. $n = 3$ has been proved by Bush [Busch, 1990] and so far[6] it is known to hold true up to $n = 5$ with a very complicated and tricky proof [Hakopian et al., 2009, Hakopian et al., 2014]. Indeed it has turned out that this problem has some very interesting deep connections to algebraic geometry [Fieldsteel and Schenck, 2017] and relies strongly on ideals and syzygies, cf. [Carnicer and Sauer, 2018].

[5] They even have a Wikipedia page.
[6] June 2019.

### 4.2.2 Constructing spaces

The opposite approach is to consider the set of sites, $\mathscr{X}$, or, more generally, the set $\Theta$ of functionals as given and to construct an appropriate unique interpolation space $\mathscr{P}$. The spaces are conveniently described by their behavior on polynomials itself, i.e., by the action of the interpolation operator $L : \Pi \to \mathscr{P}$.

The question of existence of such spaces is easily solved as long as the interpolation problem is an ideal one.

**Theorem 4.2.10.** *For each ideal interpolation problem there exists at least one unique interpolation space.*

**Proof:** Since the problem is ideal, $\mathscr{I} = \ker\Theta$ is an ideal and, due to the general assumption that $\dim\Theta < \infty$, Proposition 3.2.12 implies that $\Pi/\mathscr{I}$ is finite dimensional. For an any grading $\Gamma$ there exists a $\Gamma$ basis $G$ for $\mathscr{I}$ and the normal form space $\mathscr{P} := \nu_{\mathscr{I}}(\Pi)$ is well defined and has dimension $\#\Theta$. Let $P$ denote any basis of $\mathscr{P}$. Since any $v \in \mathbb{K}^P$ such that

$$0 = V(P,\Theta)\, v = \big(\theta(v \cdot P) : \theta \in \Theta\big) \qquad \Rightarrow \qquad v \cdot P \in (\Pi/\mathscr{I}) \cap \mathscr{I} = \{0\},$$

must satisfy $v = 0$, the Vandermonde matrix $V(P,\Theta)$ is nonsingular and therefore $\mathscr{P}$ is a unique interpolation space. □

**Definition 4.2.11.** The CANONICAL INTERPOLATION SPACE for an ideal interpolation problem with respect to $\Theta \subset \Pi'$ is $\nu_{\mathscr{I}}(\Pi)$, where $\mathscr{I} := \ker\Theta$. The FUNDAMENTAL POLYNOMIAL $\ell_\theta$ is defined as

$$\ell_\theta := v_\theta \cdot P := \big(V(P,\Theta)^{-1} e_\theta\big) \cdot P, \qquad \theta \in \Theta. \tag{4.2.3}$$

**Remark 4.2.12.** The existence of the fundamental polynomials follows from the proof of Theorem 4.2.10 where the nonsingularity of the Vandermonde matrix has been shown. The fundamental polynomials satisfy

$$\theta'(\ell_\theta) = \theta'(v_\theta \cdot P) = e_{\theta'} V_{P,\Theta}\, v_\theta = e_{\theta'}^T V_{P,\Theta} V(P,\Theta)^{-1} e_\theta = e_{\theta'}^T e_\theta = \delta_{\theta,\theta'} \tag{4.2.4}$$

as expected.

**Definition 4.2.13.** The DEGREE of a subspace $\mathscr{P} \subset \Pi$ with respect to a given grading $\Gamma$ is defined as

$$\delta(\mathscr{P}) = \sup\{\delta(p) : p \in \mathscr{P}\}. \tag{4.2.5}$$

If $\mathscr{P}$ is finite dimensional, the "sup" in (4.2.5) can be replaced by "max".

**Definition 4.2.14** (Minimal degree & degree reducing)**.** A unique interpolation space $\mathscr{P} \subset \Pi$ with respect to $\Theta \subset \Pi'$ is called

1. of MINIMAL DEGREE if any unique interpolation space $\mathscr{Q} \subset \Pi$ satisfies $\delta(\mathscr{Q}) \geq \delta(\mathscr{P})$.
2. DEGREE REDUCING if
$$\delta\big(L_\Theta f\big) \leq \delta(f), \qquad f \in \Pi, \tag{4.2.6}$$
where $L_\Theta : \Pi \to \mathscr{P}$ denotes the INTERPOLATION OPERATOR defined by
$$\theta\big(L_\Theta f\big) = \theta(f), \qquad \theta \in \Theta. \tag{4.2.7}$$

**Proposition 4.2.15.** *Any degree reducing interpolation space is of minimal degree.*

**Proof:** Let $\mathscr{P}$ be a degree reducing space, $L_\Theta : \Pi \to \mathscr{P}$ the associated interpolation operator and assume that $\mathscr{Q}$ is a unique interpolation space of smaller degree, $\delta(\mathscr{Q}) < \delta(\mathscr{P})$. Let $\ell'_\theta \in \mathscr{Q}$, $\theta \in \Theta$, be the fundamental polynomials, then, since $\mathscr{P}$ is degree reducing,

$$\delta(\ell_\theta) = \delta\left(L_\Theta \ell'_\theta\right) \le \delta\left(\ell'_\theta\right) \le \deg \mathscr{Q} < \deg(\mathscr{P}), \qquad \theta \in \Theta,$$

which means that any $f \in \mathscr{P}$ satisfies

$$\delta(f) = \delta\left(\sum_{\theta\in\Theta} \theta(f)\,\ell_\theta\right) \le \max_{\theta\in\Theta} \delta(\ell_\theta) < \deg \mathscr{P} \tag{4.2.8}$$

which is a contradiction. □

From (4.2.8) we get the following conclusion.

**Corollary 4.2.16.** *For any unique interpolation space we have that* $\max\{\delta(\ell_\theta) : \theta \in \Theta\} = \delta(\mathscr{P})$.

**Exercise 4.2.1** Show that if $\mathscr{P} = \Pi_n$ is a unique interpolation space for the sites $\mathscr{X}$ then $\deg \ell_x = n$, $x \in \mathscr{X}$. ♢

**Theorem 4.2.17.** *The canonical interpolation space is degree reducing.*

**Proof:** The fact that for any $\Gamma$ basis $G$

$$f = \sum_{g\in G} f_g\, g + \nu_{\mathscr{I}}(f)$$

is a $\Gamma$-representation implies that $\delta\left(\nu_{\mathscr{I}}(f)\right) \le \delta(f)$, hence interpolation is degree reducing. □

**Remark 4.2.18.** The canonical interpolation space $\nu_{\mathscr{I}}(\Pi)$ and its basis $\ell_\theta$ are a template for any other unique interpolation space. Indeed let $\ell'_\theta$ be the fundamental polynomials of another unique interpolation space, then $\ell_\theta - \ell_\theta \in \mathscr{I}$. In other words: *any unique interpolationspace $\mathscr{P}$ can be written as*

$$\mathscr{P} = \operatorname{span}\left\{\ell_\theta + f_\theta : f_\theta \in \mathscr{I}, \theta \in \Theta\right\} \tag{4.2.9}$$

*where $\ell_\theta$ are the fundamental polynomials of the canonical interpolation space.* Not, however, that in most cases the space in (4.2.9) will not be degree reducing any more, this will be elaborated in Section 4.3.1.

In the case of a monomial grading, that is, a Gröbner basis, the canonical interpolation space is particularly simple.

**Example 4.2.19.** If $\Gamma$ is a monomial grading, then the monomial ideal $\lambda(\mathscr{I})$ is of the form $\lambda(\mathscr{I}) = \operatorname{span}\{x^U\}$ for some upper set $U \subseteq \mathbb{N}_0^s$. Hence, the canonical interpolation space takes the form

$$\nu_{\mathscr{I}}(\Pi) = \operatorname{span}\{x^A\}, \qquad A = \mathbb{N}_0^s \setminus U. \tag{4.2.10}$$

In other words, canonical interpolation spaces with respect to monomial gradings always correspond to lower sets.

**Definition 4.2.20.** For $A \subset \mathbb{N}_0^s$, we denote by

$$\Pi_A := \operatorname{span}\{x^A\} = \left\{\sum_{\alpha\in A} f_\alpha\, (\cdot)^\alpha : f_\alpha \in \mathbb{K}\right\} \tag{4.2.11}$$

the monomial space generated by the exponents from the set $A$.

We can express the observations from Example 4.2.19 in the following way.

**Corollary 4.2.21.** *For any ideal interpolation problem based on $\Theta \subset \Pi'$ there exists a lower set $A \subset \mathbb{N}_0^s$ such that $\Pi_A$ is a unique interpolation space with respect to $\Theta$.*

## 4.3 Ideal interpolation constructions

We now take a closer look at properties of canonical interpolation spaces.

### 4.3.1 Newton bases and ideals from points

Newton bases are a useful way of describing Lagrange interpolation problems and provide a nice equivalence between the existence of such a basis and degree reduction that also allows us to create "ideals from points".

**Definition 4.3.1.** A subspace $\mathscr{Q} \subseteq \Pi$ is called HOMOGENEOUSLY GENERATED if

$$\mathscr{Q} = \bigoplus_{\gamma \in \Gamma} \left( \mathscr{Q} \cap \Pi_\gamma \right). \tag{4.3.1}$$

**Proposition 4.3.2.** *Any canonical interpolation space $\mathscr{P} \subset \Pi$ is homogeneously generated and*

$$W_\gamma(\mathscr{I}) := W_\gamma(G), \qquad \mathscr{I} = \langle G \rangle, \tag{4.3.2}$$

*depends only on $\mathscr{I}$, not on the particular $\Gamma$-basis.*

**Proof:** If $f = \sum_\gamma f_\gamma \in \mathscr{P}$, then $f_\gamma \in W_\gamma(G)$, for a $\Gamma$ basis $G$ of $\mathscr{I} = \ker\Theta$. If we apply reduction to one of these components $f_\gamma$ then its projection on $V_\gamma(G)$ is zero and therefore $\nu_{\mathscr{I}}\left(f_\gamma\right) = f_\gamma$, hence $f_\gamma \in \nu_{\mathscr{I}}(\Pi) = \mathscr{P}$. In other words,

$$W_\gamma(G) = \nu_{\mathscr{I}}\left(W_\gamma(G')\right)$$

for any two $\Gamma$-bases $G$, $G'$ of $\mathscr{I}$, hence depends only on $\mathscr{I}$. □

If $\mathscr{P}$ is a homogeneously generated space, the we can order the set of all relevant homogeneous spaces

$$\Gamma_{\mathscr{P}} = \left\{ \gamma \in \Gamma : \mathscr{P} \cap \Pi_\gamma \neq \{0\} \right\},$$

by size and write it as

$$\Gamma_{\mathscr{P}} = \left\{ \gamma^0, \ldots, \gamma^m \right\}, \qquad m = \#\Gamma_{\mathscr{P}} - 1. \tag{4.3.3}$$

For $k = 0, \ldots, m$ this defines natural subspaces

$$\mathscr{P}_k^0 = \mathscr{P} \cap \Pi_{\gamma^k}, \qquad \mathscr{P}_k = \bigoplus_{j=0}^{k} \mathscr{P}_j^0, \qquad \mathscr{P}_k^0 \subseteq \mathscr{P}_k \subseteq \mathscr{P}. \tag{4.3.4}$$

of $\mathscr{P}$. This allows us to extend the basic concept of the NEWTON APPROACH to several variables, namely interpolation by increasing degree with polynomials that vanish at "earlier" points.

**Definition 4.3.3** (Newton basis)**.** A subset

$$N = \bigcup_{k=0}^{m} N_k, \qquad N_k \subset \mathscr{P}_k,$$

of $\mathscr{P}$ is called a NEWTON BASIS provided that

1. there exists a decomposition $\mathscr{X} = \mathscr{X}_0 \cup \cdots \cup \mathscr{X}_m$ such that

$$N_k\left(\mathscr{X}_j\right) = 0, \quad 0 \leq j < k \leq m, \qquad N_k\left(\mathscr{X}_k\right) = I, \quad k = 0, \ldots, m, \tag{4.3.5}$$

2. and

$$\Pi_\gamma = \big(\lambda(N) \cap \Pi_\gamma\big) \oplus \lambda\big(\mathscr{I}(X)\big), \qquad \gamma \in \Gamma. \tag{4.3.6}$$

**Remark 4.3.4.** Condition (4.3.5) generalizes the main idea of the Newton approach, namely that

$$(\cdot - x_0)\cdots(\cdot - x_{k-1})$$

vanishes at the "earlier" points $x_0, \ldots, x_{k-1}$, we only normalized the polynomial differently to make it 1 and $x_k$ instead of monic. Condition (4.3.6), on the other hand, does *not* exist in the univariate case, it is a (necessary) ideal theoretic extension.

**Lemma 4.3.5.** *If $\mathscr{P}$ is a canonical interpolation space, then*

$$\Gamma_{\mathscr{P}} = \big\{\gamma \in \Gamma \,:\, W_\gamma(\mathscr{I}) \neq \{0\}\big\}.$$

**Proof:** For each $\gamma \in \Gamma_{\mathscr{P}}$ there exist a polynomial $0 \neq p \in \Pi_\gamma \cap \mathscr{P}$ since, by Proposition 4.3.2 $\mathscr{P}$ is homogeneously generated. Since $p = \nu_{\mathscr{I}}(p) \in W_\gamma(\mathscr{I})$ and therefore

$$\{0\} \neq \{p\} \subset \Gamma_{\mathscr{I}} := \big\{\gamma \in \Gamma \,:\, W_\gamma(\mathscr{I}) \neq \{0\}\big\},$$

we conclude that $\Gamma_{\mathscr{P}} \subseteq \Gamma_{\mathscr{I}}$. Conversely, if $\gamma \in \Gamma_{\mathscr{I}}$, there exists $0 \neq f \in W_\gamma(\mathscr{I})$ whose normal form is $f = \nu_{\mathscr{I}}(f) \in \mathscr{P}$, hence $\gamma \in \Gamma_{\mathscr{P}}$, that is $\Gamma_{\mathscr{I}} \subseteq \Gamma_{\mathscr{P}}$. □

We can now give a construction of a Newton basis that is actually a generalization or application of the Gram-Schmidt orthogonalization process.

To that end, we recall (4.3.3), begin with $\mathscr{P}_0 = \mathscr{P} \cap \Pi_{\gamma^0}$, choose a basis $P_0$ of $\mathscr{P}_0$ and form the matrix

$$P_0(\mathscr{X}) = \big(p(x) \,:\, p \in P_0,\, x \in \mathscr{X}\big) \in \mathbb{K}^{P_0 \times \mathscr{X}}.$$

If there exists $v \in \mathbb{K}^{P_0}$ such that

$$0 = v^T P_0(\mathscr{X}) = \big(\textstyle\sum_{p \in P_0} v_p\, p(x) \,:\, x \in X\big) = (v \cdot P_0)(\mathscr{X}),$$

then $v \cdot P_0 \in \mathscr{P} \cap \mathscr{I}$, hence $v = 0$. This the rows of rank $P_0(\mathscr{X}) = \#P_0$ and[7] there exists $\mathscr{X}_0 \subseteq \mathscr{X}$, $\#\mathscr{X}_0 = \dim \mathscr{P}_0$, such that $P_0(\mathscr{X}_0)$ is nonsingular. The rows of the inverse[8] $P_0(\mathscr{X}_0)^{-1}$, are coefficient vectors for polynomials, giving

$$N_0 := P_0(\mathscr{X}_0)^{-1} P_0 \tag{4.3.7}$$

as a vector of polynomials with

$$N_0(\mathscr{X}_0) = P_0(\mathscr{X}_0)^{-1} P_0(\mathscr{X}_0) = I.$$

This introduces a 1−1 relationship between $N_0$ and $\mathscr{X}_0$ and allows us to index $N_0 = (n_x \,:\, x \in \mathscr{X}_0)$ or $\mathscr{X}_0 = [x_n \,:\, n \in N_0]$, whichever we find more convenient.

The next step consists in setting $\mathscr{X}_1' = \mathscr{X} \setminus \mathscr{X}_0$ as the set of "free" points, to choose a basis $P_1'$ of $\mathscr{P}_1 := \mathscr{P} \cap \Pi_{\gamma^1}$ to make this basis vanish at $\mathscr{X}_0$ by

$$P_1 = P_1' - N_0^T P_1'(\mathscr{X}_0), \qquad \text{i.e.,} \qquad p := p' - \sum_{x \in \mathscr{X}_0} p'(x)\, n_x, \quad p' \in P_1.$$

[7] Since $\dim \mathscr{P}_0 \le \dim \mathscr{P} = \#X$ the matrix has fewer rows than columns.

[8] Note that the inverse of a matrix in $\mathbb{K}^{P_0 \times \mathscr{X}}$ belongs to $\mathbb{K}^{\mathscr{X} \times P_0}$.

Indeed,

$$P_1(\mathscr{X}_0) = P_1'(\mathscr{X}_0) - \underbrace{N_0(\mathscr{X}_0)}_{=I} P_1'(\mathscr{X}_0) = P_1'(\mathscr{X}_0) - P_1(\mathscr{X}_0) = 0$$

hence $P_1 \in \mathscr{P}_0 + \mathscr{P}_1 \subseteq \mathscr{P}$ satisfies the annihilation part of (4.3.5). With the same argument as above we again obtain that $\operatorname{rank} P_1\left(\mathscr{X}_1'\right) = \#P_1 = \dim \mathscr{P}_1$ as $v^T P_1\left(\mathscr{X}_1'\right) = 0$ yields that $v \cdot P_1$ vanishes at $\mathscr{X}_1'$ but also at $\mathscr{X}_0$ by construction of $P_1$, hence $v \cdot P_1 \in \mathscr{I}$ and therefore $v = 0$. Thus, there are points $\mathscr{X}_1 \subseteq \mathscr{X}_1'$, such that $P_1(\mathscr{X}_1)$ is invertible, yielding

$$N_1 := P_1(X_1)^{-1} P_1. \tag{4.3.8}$$

The general step with index $k$ works in exactly the same way: we set

$$\mathscr{X}_k' := \mathscr{X} \setminus \bigcup_{j=0}^{k-1} \mathscr{X}_j,$$

choose a basis $P_k'$ of $\mathscr{P}_k := \mathscr{P} \cap \Pi_{\gamma^k}$ and ensure, since the block matrix appearing in (4.3.9) has identity matrices on the diagonal,

$$P_k := P_k' - \left(P_k'(\mathscr{X}_0) \ldots P_k'(\mathscr{X}_{k-1})\right) \begin{pmatrix} N_0(\mathscr{X}_0) & \ldots & N_0(\mathscr{X}_{k-1}) \\ & \ddots & \vdots \\ & & N_{k-1}(\mathscr{X}_{k-1}) \end{pmatrix} \begin{pmatrix} N_0 \\ \vdots \\ N_{k-1} \end{pmatrix} \tag{4.3.9}$$

that

$$P_k\left(\mathscr{X}_j\right) = 0, \qquad j = 0, \ldots, k-1.$$

Since, by the meanwhile well-known argument, $\operatorname{rank} P_k(\mathscr{X}_k) = \#P_k = \dim \mathscr{P}_k$ there exists $\mathscr{X}_k \subseteq \mathscr{X}_k' \subseteq \mathscr{X}$, such that $P_k(\mathscr{X}_k)$ is nonsingular, giving

$$N_k := P_k(\mathscr{X}_k)^{-1} P_k. \tag{4.3.10}$$

**Remark 4.3.6.** This procedure can be seen as a Gram-Schmidt process, but also as a blockwise GAUSSIAN ELIMINATION applied to the Vandermonde matrix[9] $P(X)$ where $P$ is a graded basis of $\mathscr{P}$, cf. [Boor, 1994]. In terms of numerical linear algebra, (4.3.9) could also be interpreted as a BACK SUBSTITUTION.

**Remark 4.3.7.** The choice of $\mathscr{X}_k$, $k = 0, \ldots, m$, is *not* unique in general, quite the contrary, the generic case is that *any* subset of $\mathscr{X}_k'$ of proper cardinality can be chosen as $\mathscr{X}_k$. The strategy to choose these points can be seen as a PIVOTING STRATEGY, yet another concept from numerical linear algebra.

The following statement just summarized the construction, there is nothing left to prove.

**Theorem 4.3.8.** *The polynomials $N = [N_k : k = 0, \ldots, m]$ constructed above form a Newton basis of $\mathscr{P}$.*

But this is only half of the truth. In fact, the existence of a Newton basis even *characterizes* degree reducing interpolation spaces, at least as soon as the are homogeneously generated.

**Theorem 4.3.9.** *A homogeneously generated subspace $\mathscr{P} \subset \Pi$ is a degree reducing interpolation space with respect to $\mathscr{X} \subset \mathbb{K}^n$ if and only if it has a Newton basis.*

[9] To be precise: the transpose of the Vandermonde matrix.

**Proof:** The trick in the proof is to connect everything to the canonical interpolation space $\mathscr{P}^* := \nu_{\mathscr{I}}(\Pi)$ and its Newton basis, which we denote by $N^*$.
"⇒": We set $N := L_{\mathscr{P}}(N^*)$ and since $N(\mathscr{X}) = N^*(\mathscr{X})$, the property (4.3.5) is already satisfied. Writing $f \in \Pi$ as $f = g + \nu_{\mathscr{I}}(f)$, $g \in \mathscr{I} = \mathscr{I}(\mathscr{X})$, we get

$$L_{\mathscr{P}} f = L_{\mathscr{P}}\left(g + \nu_{\mathscr{I}}(f)\right) = \underbrace{L_{\mathscr{P}} g}_{=0} + L_{\mathscr{P}} \nu_{\mathscr{I}}(f) = L_{\mathscr{P}} \nu_{\mathscr{I}}(f)$$

and since $L_{\mathscr{P}}$ and the normal form operation are degree reducing, it follows that

$$\delta\left(L_{\mathscr{P}} f\right) = \delta\left(L_{\mathscr{P}} \nu_{\mathscr{I}(X)}(f)\right) \le \delta\left(\nu_{\mathscr{I}(X)}(f)\right) = \delta\left(\nu_{\mathscr{I}(X)}\left(L_{\mathscr{P}} f\right)\right) \le \delta\left(L_{\mathscr{P}} f\right),$$

hence $\delta\left(L_{\mathscr{P}} f\right) = \delta\left(\nu_{\mathscr{I}}(f)\right) =: \gamma$. Both leading terms therefore have the same degree $\gamma = \gamma_k$ and

$$q_x := \lambda(n_x) - \lambda\left(n_x^*\right) \in V_\gamma(\mathscr{I}), \qquad x \in \mathscr{X}_k.$$

This implies that

$$\begin{aligned}
\Pi_\gamma &= \operatorname{span}\left\{\lambda\left(n_x^*\right) : x \in \mathscr{X}_k\right\} + V_\gamma(\mathscr{I}) \\
&= \operatorname{span}\left\{\lambda(n_x) - q_x : x \in \mathscr{X}_k\right\} + V_\gamma(\mathscr{I}) \\
&\subseteq \operatorname{span}\left\{\lambda(n_x) : x \in \mathscr{X}_k\right\} + \underbrace{\operatorname{span}\left\{q_x : x \in \mathscr{X}_k\right\}}_{\subseteq V_\gamma(\mathscr{I})} + V_\gamma(\mathscr{I}) \\
&= \left(\lambda(N) \cap \Pi_\gamma\right) + V_\gamma(\mathscr{I}) \subseteq \Pi_\gamma,
\end{aligned}$$

hence

$$\Pi_\gamma = \left(\lambda(N) \cap \Pi_\gamma\right) + V_\gamma(\mathscr{I}),$$

which is (4.3.6).
"⇐": Since $\mathscr{P} = \operatorname{span} N$ is an interpolation space and the matrix

$$N(\mathscr{X}) = \begin{pmatrix} N_0(\mathscr{X}_0) & N_0(\mathscr{X}_1) & \dots & N_0(\mathscr{X}_m) \\ N_1(\mathscr{X}_0) & N_1(\mathscr{X}_1) & \ddots & \vdots \\ \vdots & \ddots & \ddots & N_{m-1}(\mathscr{X}_m) \\ N_m(\mathscr{X}_0) & \dots & N_m(\mathscr{X}_{m-1}) & N_m(\mathscr{X}_m) \end{pmatrix} = \begin{pmatrix} I & * & \dots & * \\ 0 & I & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \dots & 0 & I \end{pmatrix}$$

is nonsingular and upper triangular, we can write the interpolant to $f \in \Pi$ as

$$L_{\mathscr{P}} f = \left(f(\mathscr{X}_0) \dots f(\mathscr{X}_m)\right) N(\mathscr{X})^{-T} \begin{pmatrix} N_0 \\ \vdots \\ N_m \end{pmatrix}.$$

To verify degree reduction, we use (4.3.6) and set $N^* = \nu_{\mathscr{I}}(N)$. The normal forms share the interpolation properties of $N$, hence form a Newton basis and satisfy $N^*(\mathscr{X}) = N(\mathscr{X})$, so that the coefficient of $L_{\mathscr{P}} f$ and $L_{\mathscr{P}^*}$ with respect to the Newton bases are the same, namely $N(\mathscr{X})^{-1} f(X) = (N(\mathscr{X})^*)^{-1} f(\mathscr{X})$. Now (4.3.6) tells us together with

$$\lambda\left(N_\gamma^*\right) - \lambda\left(N_\gamma\right) \in V_\gamma(\mathscr{I}(X)), \qquad \Gamma \in \Gamma_{\mathscr{P}},$$

that

$$\gamma = \delta\left(N_\gamma^*\right) = \delta\left(N_\gamma^*\right), \qquad \gamma \in \Gamma_{\mathscr{P}} = \Gamma^*,$$

and since the normal form is degree reducing, $\mathscr{P}$ must be a degree reducing interpolation space. □

The final step is the construction of a $\Gamma$-basis for $\mathscr{I} = \mathscr{I}(\mathscr{X})$ from the set $\mathscr{X}$ of sites. This will repeat the construction of the Newton basis, but without knowing $\Gamma_{\mathscr{P}}$ – instead, we will determine it from the sites as well. The next result is fairly obvious from the definition of a well-ordering, nevertheless we will give a quick proof for the sake of completeness.

**Lemma 4.3.10.** *Each subset $\Gamma' \subseteq \Gamma$ of a well-ordered monoid has a smallest element.*

**Proof:** For $\gamma^0 \in \Gamma'$ we consider $\Gamma'_1 = \{\gamma \in \Gamma' : \gamma < \gamma^0\}$. If $\Gamma'_1 = \emptyset$, then $\gamma^0$ is minimal, otherwise we choose $\gamma^1 \in \Gamma'_1$, satisfying $\gamma^1 < \gamma^0$, and so on, yielding a strictly descending chain $\gamma^0 > \gamma^1 > \gamma^2 > \cdots$ end which has to be finite and yields the minimal element. □

We start the construction with $\gamma = \min\Gamma$, which has to exist according to Lemma 4.3.10. To be honest, this in not necessary in the initial step since Lemma 2.2.8 tells us that $\gamma = 0$ is the choice. The we choose a basis[10] $P$ of $\Pi_\gamma$ and consider the matrix $P(\mathscr{X})$, whose rank is between 0 and $\#X$. Even if the basis $P$ is infinite, this rank is always *finite*. This allows us to find in $P(X)$ a *square* and *nonsingular*[11] of maximal rank. Associated to this matrix are subsets $P_\gamma \subseteq P$ and $\mathscr{X}_\gamma \subseteq \mathscr{X}$, such that $P_\gamma(\mathscr{X}_\gamma)$ is a *maximal* nonsingular submatrix of $P(\mathscr{X})$ and after reordering $P$ and $\mathscr{X}$ we have, setting $\overline{P_\gamma} := P \setminus P_\gamma$ and $\overline{\mathscr{X}_\gamma} := \mathscr{X} \setminus \mathscr{X}_\gamma$ the block representation

$$P(\mathscr{X}) = \begin{pmatrix} P_\gamma(\mathscr{X}_\gamma) & P_\gamma\left(\overline{\mathscr{X}_\gamma}\right) \\ \overline{P_\gamma}(\mathscr{X}_\gamma) & \overline{P_\gamma}\left(\overline{\mathscr{X}_\gamma}\right) \end{pmatrix} = \begin{pmatrix} I & 0 \\ \overline{P_\gamma}(\mathscr{X}_\gamma)\, P_\gamma(\mathscr{X}_\gamma)^{-1} & I \end{pmatrix} \begin{pmatrix} P_\gamma(\mathscr{X}_\gamma) & P_\gamma\left(\overline{\mathscr{X}_\gamma}\right) \\ 0 & * \end{pmatrix}.$$

The "$*$" part in the matrix on the right hand side must be zero as otherwise the rank of the matrix would exceed $P_\gamma(\mathscr{X}_\gamma)$, hence also $P(\mathscr{X})$ and therefore

$$\begin{pmatrix} P_\gamma(\mathscr{X}_\gamma) & P_\gamma\left(\overline{\mathscr{X}_\gamma}\right) \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} I & 0 \\ -\overline{P_\gamma}(\mathscr{X}_\gamma)\, P_\gamma(\mathscr{X}_\gamma)^{-1} & I \end{pmatrix} \begin{pmatrix} P_\gamma(\mathscr{X}_\gamma) & P_\gamma\left(\overline{\mathscr{X}_\gamma}\right) \\ \overline{P_\gamma}(\mathscr{X}_\gamma) & \overline{P_\gamma}\left(\overline{\mathscr{X}_\gamma}\right) \end{pmatrix}. \tag{4.3.11}$$

Therefore,

$$Q_\gamma := \overline{P_\gamma} - \overline{P_\gamma}(\mathscr{X}_\gamma)\, P_\gamma(\mathscr{X}_\gamma)^{-1} P_\gamma$$

satisfies $Q_\gamma(\mathscr{X}) = 0$, hence $Q_\gamma \subset \mathscr{I}$. The construction can be summarized as follows:

1. $P_\gamma$ and $\mathscr{X}_\gamma$ admit *unique* interpolation: $\det P_\gamma(\mathscr{X}_\gamma) \neq 0$.
2. Subsets $Q_\gamma$ with $Q_\gamma(\mathscr{X}) = 0$, hence $Q_\gamma \subset \mathscr{I}$.
3. Together, these subsets generate $\Pi_\gamma$:

$$\Pi_\gamma = P_\gamma \oplus Q_\gamma \tag{4.3.12}$$

Thus, the decomposition from (4.3.12) is a very useful one: the subset $P_\gamma$ can be transformed into a part of a Newton basis $N$ by,

$$N_\gamma := P_\gamma(X_\gamma)^{-1} P_\gamma,$$

[10] This can be infinite, so a little bit of care is necessary here.
[11] For a nonsingular matrix, square is redundant.

the other subset, $Q_\gamma$, is part of an ideal basis $Q$; we add $N_\gamma$ to $N$, $Q_\gamma$ to $Q$ and replace $\mathscr{X}$ by $\mathscr{X}' = \mathscr{X} \setminus \mathscr{X}_\gamma$, just as in the previous construction for the Newton basis for the canonical interpolation space.

The set $Q$ consists of polynomials that vanish at $\mathscr{X}$ so that $\mathscr{I} = \mathscr{I}(X) \subseteq \langle Q \rangle$, in particular, we do not have to search for interpolation polynomials any more in the spaces $V_\eta(Q)$, $\eta > \gamma$, as those are leading parts of polynomials in the ideal. The next degree to check is therefore

$$\gamma' := \min\{\eta \geq \gamma : W_\eta(Q) \neq \{0\}\},$$

which exists by Lemma 4.3.10. We choose $P$ as a basis of $W_{\gamma'}(Q)$ and by subtracting interpolants with respect to the our Newton basis built so far, we can again ensure that $P(X_\eta) = 0$, $\eta \leq \gamma$, without changing the degree of any element of $P$ as the homogeneous terms we started with have degree $\gamma' > \gamma$. The next step is to decompose the $P(X')$ by means of kernel and range into ideal and interpolation polynomials. In such a step it may well happen that either $P_{\gamma'} = \mathscr{X}_{\gamma'} = \emptyset$ or $Q_{\gamma'} = \emptyset$, the cases being mutually exclusive since $P_{\gamma'}$ and $Q_{\gamma'}$ generate the *nontrivial* vector space $W_{\gamma'}(Q)$. After the update step

$$\begin{aligned} N &= N \cup P_{\gamma'} \left(\mathscr{X}_{\gamma'}\right)^{-1} P_{\gamma'} && (4.3.13)\\ Q &= Q \cup Q_{\gamma'} && (4.3.14)\\ \mathscr{X}' &= \mathscr{X}' \setminus \mathscr{X}_{\gamma'} && (4.3.15) \end{aligned}$$

we continue with the iteration with the effect that in each step either $\mathscr{X}'$ is strictly reduced or $\langle \lambda(Q) \rangle_h$ is strictly enlarged, maybe even both. Since this can be done only *finitely many* times, the procedure terminates with as final $\gamma \in \Gamma$ such that

$$\Pi_\eta = V_\eta(Q), \qquad \eta > \gamma,$$

By construction we also have that

$$\lambda(\mathscr{I}) \cap \left( \bigoplus_{\eta \leq \gamma} \Pi_\eta \right) \subset \bigoplus_{\eta \leq \gamma} V_\eta(Q),$$

since each polynomial from the ideal whose leading part does not belong to $V_\gamma(Q)$ for some $\gamma \in \Gamma$ has been explicitly added to the ideal. Together with $Q \subset \mathscr{I}(X)$ and taking into account that $\langle \lambda(Q) \rangle_h = \bigoplus_{\gamma \in \Gamma} V_\gamma(Q)$ as well as Lemma 2.4.14, we can conclude that

$$\lambda(\mathscr{I}(X)) \subseteq V_\gamma(Q) \subseteq \lambda(\mathscr{I}(X)),$$

which proves our final result.

**Theorem 4.3.11.** *The set $Q \subset \mathscr{I}(X)$ is a $\Gamma$-basis for the ideal $\mathscr{I} = \mathscr{I}(\mathscr{X})$.*

### 4.3.2 Least interpolation

There is an almost explicit way to give the interpolation space for ideal interpolation problems due to Carl de Boor and Amos Ron [Boor and Ron, 1990, Boor and Ron, 1992] which was originally motivated by ideas from box spline theory [Boor and Ron, 1991]. For that purpose, we need some more notation and one assumption.

**Definition 4.3.12** (Least part)**.** Let

$$\mathbb{K}[\![x]\!] := \left\{ f(x) = \sum_{\alpha \in \mathbb{N}_0^s} f_\alpha \, x^\alpha : f_\alpha \in \mathbb{K} \right\} \tag{4.3.16}$$

denote the ring of all FORMAL POWER SERIES. Given a grading $\Gamma$ and $f \in \mathbb{K}(x)$, with homogeneous decomposition $f = \sum f_\gamma$, we denote by

$$\delta_\downarrow(f) := \max\{\gamma : f_\alpha = 0, \alpha < \gamma\} \tag{4.3.17}$$

the LEAST DEGREE of $f$ and by

$$\lambda_\downarrow(f) := f_{\delta_\downarrow(f)}$$

the MINIMAL FORM or LEAST PART of $f$. Given a subspace $\mathscr{F} \subset \mathbb{K}(x)$, we denote by

$$\lambda_\downarrow(\mathscr{F}) := \{\lambda_\downarrow(f) : f \in \mathscr{F}\} \subseteq \Pi \tag{4.3.18}$$

the subspace of all least parts in $\mathscr{F}$.

**Remark 4.3.13.** Since formal power series usually have no maximal or leading part, the only chance is to look at the origin in order to get a finite quantity.

**Definition 4.3.14.** The grading $\Gamma$ is said to be COMPATIBLE with the inner product $(f,g) = (g(D)f)(0)$ if $(\Pi_\gamma, \Pi_{\gamma'}) = 0$ for $\gamma \neq \gamma'$.

Examples for compatible gradings are all monomial gradings as well as the standard homogeneous grading. In fact, the approach from [Boor and Ron, 1992] is only considering the homogeneous grading the simplest nontrivial case where these ideas become relevant. The extension to compatible gradings is straightforward.
The following result, due to de Boor and Ron, [Boor and Ron, 1990], is based on the duality between polynomials and exponential polynomials defining the functionals of an ideal interpolation problem, according to 3.2.9.

**Theorem 4.3.15** (Least interpolation)**.** *Let an ideal interpolation problem be given by sites $\mathscr{X} \subset \mathbb{K}^s$ and $D$-invariant multiplicity spaces $\mathscr{Q}_x$ with bases $Q_x$, $x \in \mathscr{X}$. Then*

$$\lambda_\downarrow(\mathscr{F}), \qquad \mathscr{F} = \operatorname{span}\{q e_x : q \in Q_x, x \in \mathscr{X}\} \tag{4.3.19}$$

*is a degree reducing unique interpolation space.*

**Proof:** We first recall that $\mathscr{I} = \ker(\cdot, \mathscr{F})$ as used in the proof of Theorem 3.2.9. For $\gamma \in \Gamma$ we define

$$\mathscr{F}_\gamma := \lambda_\downarrow(\mathscr{F}) \cap \Pi_\gamma$$

and note that for $g \in \mathscr{I}$, $\delta(g) \le \gamma$, and $f \in \mathscr{F}$ with $\lambda_\downarrow(f) \in \mathscr{F}_\gamma$ we have, by the assumption that the grading is compatible, that

$$0 = (g,f) = \left( \sum_{\gamma' \le \gamma} g_{\gamma'}, \sum_{\gamma' \ge \gamma} f_{\gamma'} \right) = (g_\gamma, f_\gamma) = \begin{cases} (\lambda(g), \lambda_\downarrow(f)), & \delta(g) = \gamma, \\ 0, & \delta(g) < \gamma. \end{cases}$$

Therefore,

$$\mathscr{F}_\gamma \oplus (\lambda(\mathscr{I}) \cap \Pi_\gamma) = \Pi_\gamma,$$

and

$$\operatorname{rank}\left((\mathscr{F}_\gamma, q e_x) : q \in Q_x, x \in \mathscr{X}\right) = \operatorname{rank}(\mathscr{F}_\gamma, \mathscr{F}_\gamma) = \dim \mathscr{F}_\gamma,$$

which is exactly the decomposition of (4.3.12), so that the nontrivial spaces $\mathscr{F}_\gamma$ even determine the homogeneous parts for a Newton basis. □

### 4.3.3 Interpolation on grids

The simplest, but not oldest way[12] to construct multivariate interpolation problems is to consider the tensor product case as many, but not all, things have a very univariate flavor then.

**Definition 4.3.16** (Grid)**.** For finite sets $\mathscr{X}_j = \{x_{j,k} : 0 \le k < \#\mathscr{X}_j\} \subset \mathbb{K}$, $j = 1,\dots,s$, and a lower set $A \subset \mathbb{N}_0^s$, the GRID $\mathscr{X}^A$ is defined as

$$\mathscr{X}^A := \{x_\alpha := (x_{1,\alpha_1},\dots,x_{s,\alpha_s}) : \alpha \in A\}. \tag{4.3.20}$$

**Remark 4.3.17.** Since the univariate point sets are finite, the lower set $A$ must also be finite which is the standing assumption whenever we speak of a grid.

**Example 4.3.18** (Rectangular and triangular grids)**.** The two most prominent examples of grids, considered already in [Isaacson and Keller, 1966] are the following:

1. The RECTANGULAR GRID uses the hypercube

$$A = [0,(\#\mathscr{X}_1)\times\cdots\times(\#\mathscr{X}_s)] := \{\alpha \in \mathbb{Z}^s : 0 \le \alpha_j \le \#\mathscr{X}_j,\ j = 1,\dots,s\} \tag{4.3.21}$$

   as index set.

2. The TRIANGULAR GRID of order $n$ uses

$$A = \{\alpha \in \mathbb{N}_0^s : |\alpha| \le n\}, \qquad n \le \min_{j=1,\dots,s} \#\mathscr{X}_j. \tag{4.3.22}$$

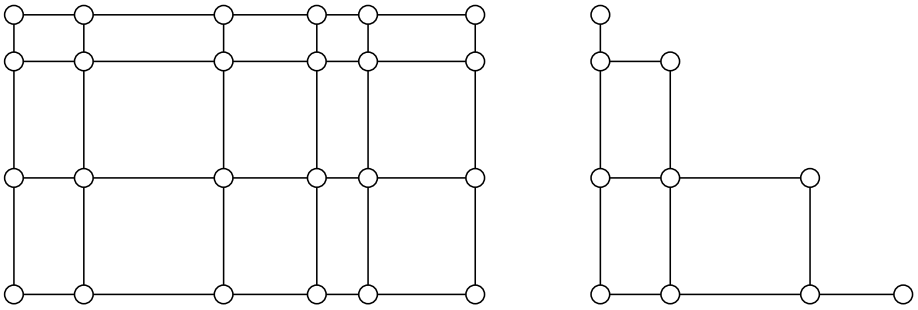


Figure 4.3.3: A rectangular *(left)* and a triangular *(right)* grid extracted from the same univariate distribution of points. Not that along the diagonals the points of the triangular grid to *not* have to lie on a straight line. Configurations of the type are the so-called PRINCIPAL LATTICES introduced in [Lee and Phillips, 1988]; they also belong to the context of the geometric characterization from Definition 4.2.5.

It is not hard to guess an interpolation space for grids. The result itself is a consequence of Theorem 4.3.23 that even shows that $\Pi_A$ is *the* canonical interpolation space for $\mathscr{X}^A$.

**Corollary 4.3.19.** *For any grid $\mathscr{X}^A$ the space $\Pi_A$ is a unique interpolation space.*

More fascinating is the fact from [Sauer, 2004] which can even be seen as a sort of converse to Corollary 4.3.19, namely that any canonical interpolation space even *must* be $\Pi_A$ under quite mild assumptions. Let us recall their definition.

[12] This is [Jacobi, 1835, Borchardt, 1860, Kronecker, 1866], dealing with so-called COMPLETE INTERSECTION.

**Definition 4.3.20.** A grading $\Gamma$ is called a MONOMIAL GRADING if $\Pi_\gamma$ is spanned by monomials as a $\mathbb{K}$-vector space and it is called a STRICT GRADING if $\delta(f) = 0$ implies that $f$ is a constant polynomial, see also Definition 2.2.13.

**Remark 4.3.21.** Any strict monomial grading can be refined to a term order by

$$(\cdot)^\alpha \prec (\cdot)^\beta \quad \Leftrightarrow \quad \delta\left((\cdot)^\alpha\right) < \delta\left((\cdot)^\alpha\right) \quad \text{or} \quad \delta\left((\cdot)^\alpha\right) = \delta\left((\cdot)^\alpha\right), (\cdot)^\alpha \prec^* (\cdot)^\beta \tag{4.3.23}$$

where $\prec^*$ is an arbitrary term order, for example the lexicographic one. This is indeed the procedure that refines the homogeneous grading into gradlex.

**Definition 4.3.22** (Universal basis)**.** A finite set $G \subset \mathscr{I}$ is called a UNIVERSAL $\Gamma$-BASIS or UNIVERSAL BASIS, for short, if it is a $\Gamma$-basis for each strict monomial grading.

In other words, the existence of a universal basis tells us that the grading is in fact irrelevant for the basis. This carries over to some interpolation problems.

**Theorem 4.3.23.** *If $\Gamma$ is a strict monomial grading and $A \subset \mathbb{N}_0^s$ a lower set, then $\Pi_A$ is the canonical interpolation space of $\mathscr{X}^A$.*

To prove Theorem 4.3.23, we consider the upper set $U = \mathbb{N}_0^s \setminus A$ and its minimal generating set $G(U)$ from (2.3.9). The crucial observation is as follows.

**Proposition 4.3.24.** *The polynomials*

$$f_\alpha := \prod_{j=1}^{s} \prod_{k=0}^{\alpha_j - 1} \left((\cdot)_j - x_{j,k}\right), \qquad \alpha \in G(U), \tag{4.3.24}$$

*are a universal $\Gamma$-basis for $\mathscr{I} = \mathscr{I}(\mathscr{X})$.*

**Proof:** We first note that $f_\alpha = (\cdot)^\alpha + q$ where all powers of monomials in $q$ belong to $L(\alpha)\setminus\{\alpha\} \subset A$, hence

$$f_\alpha \in (\cdot)^\alpha + \Pi_A, \qquad \alpha \in G(U). \tag{4.3.25}$$

To see that $f_\alpha \in \mathscr{I}$, we choose any $\beta \in A$, hence $\beta \notin U(\alpha)$ so that there must exist at least one index $j$ with $\beta_j < \alpha_j$ and thus

$$\prod_{k=0}^{\alpha_j - 1} \left((x_\beta)_j - x_{j,k}\right) = \prod_{k=0}^{\alpha_j - 1} \left(x_{j,\beta} - x_{j,k}\right) = 0. \tag{4.3.26}$$

Hence $\left\langle f_\alpha : \alpha \in G(U)\right\rangle \subseteq \mathscr{I}$.

Next, we refine the grading $\Gamma$ into a term order $\Gamma^*$ as pointed out in Remark 4.3.21 and continue with $\Gamma^*$, showing that the $f_\alpha$ form a Gröbner basis with respect to the term order $\Gamma^*$. Since the grading still is strict, we have $\lambda\left((\cdot)_j - x_{j,k}\right) = (\cdot)_j$ and thus $\lambda(f_\alpha) = (\cdot)^\alpha$. For $\alpha, \alpha' \in G(U)$ set $\eta := \max(\alpha, \alpha')$, i.e., $U(\alpha) \cap U(\alpha') = U(\eta)$. With arbitrary values[13] $x_{j,k}$, $k \geq \#\mathscr{X}_j$, $j = 1, \ldots, s$, we define

$$g_\alpha := \prod_{j=1}^{s} \prod_{k=\alpha_j}^{\eta_j} \left((\cdot)_j - x_{j,k}\right), \qquad g_{\alpha'} := \prod_{j=1}^{s} \prod_{k=\alpha'_j}^{\eta_j} \left((\cdot)_j - x_{j,k}\right), \tag{4.3.27}$$

[13] We are simply adding sites that are irrelevant for $\mathscr{X}^A$, hence change nothing with the interpolation problem.

and note that $\lambda(g_\alpha) = (\cdot)^{\eta-\alpha}$ as well as $\lambda(g_{\alpha'}) = (\cdot)^{\eta-\alpha'}$ and

$$g_\alpha f_\alpha = g_{\alpha'} f_{\alpha'} = \prod_{j=1}^{s} \prod_{k=0}^{\eta_j} \left((\cdot)_j - x_{j,k}\right).$$

Consequently, we can write the S-polynomials for $\alpha, \alpha'$ as

$$\begin{aligned} s\left(f_\alpha, f_{\alpha'}\right) &= (\cdot)^{\eta-\alpha} f_\alpha - (\cdot)^{\eta-\alpha'} f_{\alpha'} \\ &= \left((\cdot)^{\eta-\alpha} - g_\alpha\right) f_\alpha - \left((\cdot)^{\eta-\alpha'} - g_{\alpha'}\right) f_{\alpha'} = \left(\lambda(g_\alpha) - g_\alpha\right) f_\alpha - \left(\lambda(g_{\alpha'}) - g_{\alpha'}\right) f_{\alpha'}, \end{aligned}$$

and since $\delta\left(\lambda(g_\alpha) - g_\alpha\right) < \delta(g_\alpha) = \delta\left((\cdot)^\alpha\right)$ and the same for $\alpha'$, any S-polynomial for two polynomials reduces to zero by means of these two polynomials. By Lemma 2.4.7, this means that $F = \left\{f_\alpha : \alpha \in G(U)\right\}$ is a Gröbner basis with respect to the grading $\Gamma^*$ regardless of the original $\Gamma$ and the refinement, hence a UNIVERSAL GRÖBNER BASIS.

By (4.3.25), $\Pi/\left\langle f_\alpha : \alpha \in G(U)\right\rangle = \Pi_A$ and since $\left\langle f_\alpha : \alpha \in G(U)\right\rangle \subseteq \mathscr{I}$, this yields that $\Pi/\mathscr{I} \subseteq \Pi_A$ and $\dim \Pi/\mathscr{I} = \#\mathscr{X}^A = \#A = \dim \Pi_A$ finally implies that the two quotient spaces and thus also the ideals coincide. □

**Proof of Theorem 4.3.23:** Any canonical interpolation space is of the form $\nu_G(\Pi)$ for some $\Gamma$-basis of $\mathscr{I}$. But since $\left\{f_\alpha : \alpha \in G(U)\right\}$ is such a $\Gamma$-basis and the quotient space does not depend on the concrete choice of the basis by Theorem 2.3.30, it can only be $\Pi_A = \nu_{\{f_\alpha\}}(\Pi)$. □

**Corollary 4.3.25.** *For the rectangular grids the only canonical interpolation space is the related space of tensor product polynomials.*

### 4.3.4 Universal interpolation

We close the chapter on interpolation by giving a partial answer to the following question:

> *Given a number $N$, what is the dimension of a space $\mathscr{P} \subset \Pi$ such that $\mathscr{P}$ is an interpolation space for any $\mathscr{X} \subset \mathbb{K}^s$, $\#\mathscr{X} = N$.*

To my knowledge[14] this problem is still unsolved. Nevertheless, these magic spaces deserve to be named.

**Definition 4.3.26.** $\mathscr{P} \subset \Pi$ is called a UNIVERSAL INTERPOLATION SPACE of order $N$ if it is an interpolation space for any $\mathscr{X} \subset \Pi$, $\#\mathscr{X} = N$.

**Remark 4.3.27** (Universal interpolation)**.**

1. By Theorem 4.1.13 any universal interpolation space in $s \geq 2$ variables must satisfy $\dim \mathscr{P} > N$, hence there exists no universal unique interpolation space as that would be a Haar space.

2. $\Pi_{N-1}$ is a universal interpolation space of order $N$ as it contains the fundamental polynomials from (4.1.1).

The dimension of the $\Pi_{N-1}$ is $\binom{N+s}{s} \approx N^s$ and thus grows polynomially in $N$ but exponentially in $s$. There exists, however, a smaller universal interpolation space.

[14] As in June 2019

**Definition 4.3.28.** The HYPERBOLIC CROSS $\Upsilon_n \subset \mathbb{Z}^s$ of degree $n$ is defined as

$$\Upsilon_n := \left\{ \alpha \in \mathbb{Z}^s : \prod_{j=1}^{s} \left(1 + |\alpha_j|\right) \le n \right\}, \qquad n \in \mathbb{N}_0. \tag{4.3.28}$$

For its POSITIVE OCTANT we write $\Upsilon_n^+ := \Upsilon_n \cap \mathbb{N}_0^s$.

The positive octant $\Upsilon_N^+$ of the hyperbolic cross is significantly smaller than $\{\alpha : |\alpha| \le N-1\} \supset \Upsilon_N^+$ since

$$\#\Upsilon_n^+ \le n \left(\log n\right)^{s-1}, \tag{4.3.29}$$

cf. [Lubich, 2008]. Despite of that it is sufficient for interpolation.

**Theorem 4.3.29.** $\Pi_{\Upsilon_N^+}$ *is a degree reducing universal interpolation space of order* $N$.

The proof of Theorem 4.3.29 consists of two simple observations that are of some independent interest.

**Definition 4.3.30.** Denote by $\mathscr{L}$ the set of all lower sets in $\mathbb{N}_0^s$ and by

$$\mathscr{L}_n := \left\{ A \in \mathscr{L} : \#A = n \right\}, \qquad n \in \mathbb{N}, \tag{4.3.30}$$

the set of all lower sets of cardinality $n$.

In what follows we consider degree reduction with respect to the total degree.

**Lemma 4.3.31.** *The space* $\Pi_{A_N}$ *is a degree reducing universal interpolation space for as ideal interpolation problems with respect* $\Theta \in \Pi'$, $\#\Theta \le N$, *where*

$$A_N := \bigcup_{j=1}^{N} \bigcup_{A \in \mathscr{L}_j} A \tag{4.3.31}$$

*is the union of all lower sets of cardinality* $\le N$.

**Proof:** Let $\mathscr{I} = \ker\Theta$, choose a Gröbner basis with respect to an arbitrary term order, then the associated degree reducing canonical interpolation space is of the form $v_{\mathscr{I}}(\Pi) = \Pi_A$ for some $A \in \mathscr{L}_{\#\Theta}$. The union of all such sets is then defines a universal interpolation space. □

**Lemma 4.3.32.** *The set* $A_N$ *from* (4.3.31) *is the positive octant of the hyperbolic cross, i.e.,*

$$\bigcup_{j=1}^{N} \bigcup_{A \in \mathscr{L}_j} A = \Upsilon_N, \qquad N \in \mathbb{N}_0. \tag{4.3.32}$$

**Proof:** Any $\alpha \in A_N$ belongs to some lower set $A \supseteq L(\alpha)$, hence

$$(1+\alpha_1)\cdots(1+\alpha_s) = \#L(\alpha) \le \#A \le N,$$

so that $A_N \subseteq \Upsilon_N$. Conversely, if $\alpha \in \Upsilon_N$, then there exists, by the same argument, $L(\alpha)$ is a lower set of cardinality $\le N$ that contains $\alpha$, hence $\alpha \in A_N$ and, consequently, $\Upsilon_N \subseteq A_N$. Together, these two inclusion yield the claim. □

Theorem 4.3.29 can be improved in the sense that the hyperbolic cross is even the *minimal* universal interpolation space. More precisely, we can state the following result.

**Theorem 4.3.33.** *If $A \subset \mathbb{N}_0^s$ defines a degree reducing universal interpolation space $\Pi_A$ of order $N$, then $\Upsilon_N^+ \subseteq A$. Hence the hyperbolic cross is the unique minimal degree reducing universal interpolation space spanned by monomials.*

Again, the proof is based on two intermediate results that make explicit use of the homogeneous grading.

**Lemma 4.3.34.** *If $A \subset \mathbb{N}_0^s$ induces a degree reducing interpolation space $\Pi_A$ with respect to a set $\mathscr{X} \subset \mathbb{K}^s$ with interpolation operator $L : \Pi \to \Pi_A$, then*

$$H := \left\{ h_\alpha = (\cdot)^\alpha - L(\cdot)^\alpha : \alpha \in \partial A \right\} \tag{4.3.33}$$

*is an H-basis of $\mathscr{I} := \mathscr{I}(\mathscr{X})$.*

**Proof:** If we write $q_\alpha := (\cdot)^\alpha - L(\cdot)^\alpha$, $\alpha \in \mathbb{N}_0^s$, then $q_\alpha = 0$, $\alpha \in A$, since $A$ is an interpolation space and $\deg q_\alpha \le |\alpha|$ since $\alpha$ is degree reducing. Moreover, $x^A$ and $Q = \{q_\alpha : \alpha \notin A\}$ form a basis of $\Pi$. Hence, any $f$ has a H-representation

$$f = Lf + (f - Lf) = Lf + \sum_{\alpha \notin A} f_\alpha \, q_\alpha, \qquad f_\alpha \in \mathbb{K},$$

with respect to the infinite H-basis $Q$. Moreover, $Q_{n+1} := Q \cap \Pi_{n+1}$, $n := \deg \Pi_A$, is a *finite* H-basis for $\mathscr{I}$ as $\Pi_{n+1}^0 \subseteq \lambda(Q_{n+1})$. Now we fix $\alpha \in \mathbb{N}_0^s$, $j \in \{1, \dots, s\}$, write

$$L(\cdot)^\alpha = \sum_{\beta \in A} c_\beta (\cdot)^\beta$$

and note that

$$\begin{aligned}
q_{\alpha+\epsilon_j}(x) - x_j q_\alpha(x) &= x^{\alpha+\epsilon_j} - L(\cdot)^{\alpha+\epsilon_j}(x) - x^{\alpha+\epsilon_j} + x_j L(\cdot)^\alpha(x) \\
&= x_j L(\cdot)^\alpha(x) - L(\cdot)^{\alpha+\epsilon_j}(x) = \sum_{\beta \in A} c_\beta x^{\beta+\epsilon_j} - L(\cdot)^{\alpha+\epsilon_j}(x) \\
&= \sum_{\beta \in \partial A} c_{\beta-\epsilon_j} x^\beta + \sum_{\beta \in A \setminus \partial A} c_{\beta-\epsilon_j} x^\beta - L(\cdot)^{\alpha+\epsilon_j}(x) \\
&= \sum_{\beta \in \partial A} c_{\beta-\epsilon_j} q_\beta(x) + \sum_{\beta \in \partial A} c_{\beta-\epsilon_j} L(\cdot)^\beta(x) + \sum_{\beta \in A \setminus \partial A} c_{\beta-\epsilon_j} x^\beta - L_A(\cdot)^{\alpha+\epsilon_j}(x) \\
&= \sum_{\beta \in \partial A} c_{\beta-\epsilon_j} q_\beta(x) + p(x), \qquad p \in \Pi_A,
\end{aligned}$$

Since $q_{\alpha+\epsilon_j}$ and $(\cdot)_j q_\alpha$ belong to $\mathscr{I}$, hence vanish at $\mathscr{X}$, it follows that

$$p(\mathscr{X}) = q_{\alpha+\epsilon_j}(\mathscr{X}) - \left((\cdot)_j q_\alpha\right)(\mathscr{X}) - \sum_{\beta \in \partial A} c_{\beta-\epsilon_j} q_\beta(\mathscr{X}) = 0,$$

hence $p \in \mathscr{I} \cap \Pi_A$ and thus $p = 0$, so that, replacing $\alpha$ by $\alpha - \epsilon_j$,

$$q_\alpha \in (\cdot)_j q_{\alpha-\epsilon_j} + \left\langle q_\alpha : \alpha \in \partial A \right\rangle \tag{4.3.34}$$

as long as $\alpha - \epsilon_j$ is defined, $\alpha \in \mathbb{N}_0^s \setminus A$, i.e., $\alpha \notin G\left(\mathbb{N}_0^s \setminus A\right)$, cf. Theorem 2.3.7. Any such $q_\alpha$ can such be reduced until $\alpha \in \partial A$. □

An inspection of the proof shows that we can replace $\partial A$ even by $G\left(\mathbb{N}_0^s \setminus A\right)$.

**Corollary 4.3.35.** *Under the assumptions of Lemma 4.3.34,*

$$H := \left\{ h_\alpha = (\cdot)^\alpha - L(\cdot)^\alpha : \alpha \in G\left(\mathbb{N}_0^s \setminus A\right) \right\} \tag{4.3.35}$$

*is a nonredundant H-basis of* $\mathscr{I} := \mathscr{I}(\mathscr{X})$.

The second result shows that any $\Pi_A$ can be seen as a normal form interpolation space as soon as it is degree reducing.

**Lemma 4.3.36.** *If* $A \subset \mathbb{N}_0^s$ *induces a degree reducing interpolation space, then there exists an inner product* $(\cdot,\cdot)_A : \mathbb{K} \to \mathbb{K}$ *such that* $\Pi_A = \nu_{\mathscr{I}}(\Pi)$ *with respect to a reduction algorithm based on* $(\cdot,\cdot)_A$.

**Proof:** We construct the inner product by setting $(\Pi_n, \Pi_{n'})_A = 0$, $n \neq n'$ and also

$$\left((\cdot)^\alpha, (\cdot)^\beta\right)_A := \left((\cdot)^\alpha, (\cdot)^\beta\right), \qquad |\alpha| = |\beta| = n, \quad \begin{cases} n < \min\{|\alpha| : \alpha \in \mathbb{N}_0^s \setminus A\}, \\ n > \deg A. \end{cases}$$

In the remaining cases use the basis $H$ of Theorem 4.3.33, and note that

$$\Pi_n^0 = (V_n(\langle H \rangle)) + \operatorname{span}\left\{x^{B_n}\right\}, \qquad B_n := \{\beta \in \mathbb{N}_0^s \setminus A : |\beta| = n\}$$

Let $H_n$ denote any basis of $V_n(\langle H \rangle)$, arrange this basis with $x^{B_n}$ into a basis $Q_n$ of $\Pi_n^0$ and compute the Gramian $G = (Q_n, Q_n^H)$ which is a hermitian and positive definite matrix, hence can be written as $G = YY^H$. We define

$$(f,g)_A := (Y^{-1}f, Y^{-1}g), \qquad f,g \in \Pi_n^0,$$

by means of the coefficient vectors of $f$ and $g$. Then

$$(Q_n, Q_n^H)_A = (Y^{-1}Q_n, Q_n^H Y^H) = Y^{-1}(Q_n, Q_n)Y^{-H} = Y^{-1}GY^{-H} = Y^{-1}\left(YY^H\right)Y^{-H} = I,$$

and in particular $\left(V_n(\langle H \rangle), (\cdot)^{B_n}\right) = 0$, hence $W_n(\langle H \rangle) = \operatorname{span}\{x^{B_n}\}$ and therefore $\Pi_A = \nu_{\mathscr{I}}(\Pi)$. □

**Proof of Theorem 4.3.33:** Let $B \subset \mathbb{N}_0^s$ be any lower set with $\#B = N$ and consider $B$ as an interpolation grid. By assumption on $A$ there exists $B' \subseteq A$ such that $\Pi_{B'}$ is a degree reducing unique interpolation space for $\mathscr{X} = B$. By Lemma 4.3.34, the polynomials $(\cdot)^\beta - L(\cdot)^\beta$, $\beta \in \partial B'$, form a Gröbner basis for $\mathscr{I}(B)$ and $\Pi_{B'} = \nu_{\mathscr{I}}(\Pi)$ with respect to the coefficient inner product. On the other hand, there exists an inner product such that $\Pi_B$ is normal form interpolation space with respect to this product and since normal form spaces for for grid interpolation are unique by Theorem 4.3.23, it follows that $B' = B$, hence $B \subseteq A$ for any lower set $B$ such that $\#B = N$. □

# Signal Processing | 5

*There are things that are facts, in a statistical sense, on paper, on a tape recorder, in evidence. And there are things that are facts because they have to be facts, because nothing makes any sense otherwise.*

(R. Chandler, *Playback*)

Now we get to the application of ideals in signal processing. Especially, we will re-use some of the concepts of interpolation, but in particular the concept of a zero and its multiples will help us in understanding properties of filters.

## 5.1 Signal spaces and filters

**Definition 5.1.1** (Signal spaces)**.** A SIGNAL $c$ is a function from $\mathbb{Z}^s \to \mathbb{R}$ and the vector space of all doubly infinite signals is denoted by $\ell(\mathbb{Z}^s)$. Moreover,

1. by $\ell_p(\mathbb{Z}^s)$, $0 \le p \le \infty$, we denote the vector spaces of all signals for which the $p$-NORM

$$\|c\|_p := \left( \sum_{\alpha \in \mathbb{Z}^s} |c_\alpha|^p \right)^{1/p}, \qquad 0 < p < \infty \tag{5.1.1}$$

   is finite, with the extension

$$\|c\|_0 := \#\text{supp}\,(c) := \#\{\alpha : c(\alpha) \neq 0\}, \qquad \|c\|_\infty := \sup_{\alpha \in \mathbb{Z}^s} |c_\alpha|. \tag{5.1.2}$$

2. the PULSE SIGNAL or simply PULSE $\delta \in \ell(\mathbb{Z}^s)$ is defined as $\delta(\alpha) = \delta_{\alpha,\beta}$.

3. the $j$th PARTIAL SHIFT OPERATOR $\tau_j$, $j = 1, \dots, s$, is defined as $\tau_j c := c(\cdot + \epsilon_j)$, and its powers as $\tau^\alpha = \tau_1^{\alpha_1} \cdots \tau_s^{\alpha_s}$, that is, $\tau^\alpha c = c(\cdot + \alpha)$, $\alpha \in \mathbb{Z}^s$.

4. the ONE-SIDED SIGNAL SPACE is defined as $\ell(\mathbb{N}_0^s) = \{c : \mathbb{N}_0^s \to \mathbb{R}\}$ with the canonical extensions to $\ell_p(\mathbb{N}_0^s)$.

5. the PARTIAL DIFFERENCE OPERATOR $\Delta^\alpha$ has the form $\Delta^\alpha = (\tau - I)^\alpha$ and for $q \in \Pi$ the DIFFERENCE OPERATOR $q(\tau)$ is defined as

$$q(\tau) = \sum_{\alpha \in \mathbb{N}_0^s} q_\alpha \tau^\alpha, \qquad q = \sum_{\alpha \in \mathbb{N}_0^s} q_\alpha (\cdot)^\alpha. \tag{5.1.3}$$

**Exercise 5.1.1** Prove the formula

$$\Delta^\alpha = (-1)^{|\alpha|} \sum_{\beta \le \alpha} (-1)^{|\beta|} \binom{\alpha}{\beta} \tau^\beta, \qquad \alpha \in \mathbb{N}_0^s, \tag{5.1.4}$$

where $\binom{\alpha}{\beta} := \binom{\alpha_1}{\beta_1} \cdots \binom{\alpha_s}{\beta_s}$ and $\beta \le \alpha$ means that $\beta_j \le \alpha_j$, $j = 1, \dots, s$. ♢

**Remark 5.1.2.** For $0 \le p < 1$ the expression $\|c\|_p$ is *not* a norm any more since it is not convex and violates the triangle inequality, which makes optimization quite difficult. Nevertheless, they have some importance and one speaks of a QUASI NORM is such cases. Here we will only use $\|\cdot\|_0$ and also that for formal convenience only.

**Remark 5.1.3.** The more natural extension of $q(D)$ would be a difference operator $q(\Delta)$ but since

$$\begin{aligned} q(\Delta) &= \sum_{|\alpha|\le \deg q} q_\alpha \Delta^\alpha = \sum_{|\alpha|\le \deg q} q_\alpha (-1)^{|\alpha|} \sum_{\beta\le\alpha} (-1)^{|\beta|} \binom{\alpha}{\beta} \tau^\beta \\ &= \sum_{|\alpha|\le \deg q} \tau^\beta (-1)^{|\beta|} \sum_{\alpha\ge\beta} (-1)^{|\alpha|} \binom{\alpha}{\beta} q_\alpha =: \sum_{|\alpha|\le \deg q} \tilde{q}_\beta \tau^\beta = \tilde{q}(\tau), \end{aligned}$$

any such operator can be expressed in terms of $q(\tau)$.

**Exercise 5.1.2** Can every difference operator of the form $q(\tau)$ also be written as $\tilde{q}(\Delta)$? ♢

**Definition 5.1.4** (Filter)**.**

1. A FILTER, more precisely an LTI FILTER (**L**inear **T**ime **I**nvariant), $F$ is a linear operator $F : \ell(\mathbb{Z}^s) \to \ell(\mathbb{Z}^s)$ that commutes with translation:

$$\tau^\alpha F = F\tau^\alpha, \qquad \alpha \in \mathbb{Z}^s. \tag{5.1.5}$$

2. The IMPULSE RESPONSE $f$ of a filter $F$ is defined as $f := F\delta \in \ell(\mathbb{Z}^s)$.

3. An FIR FILTER (**F**inite **I**mpulse **R**esponse) is a filter with finitely supported impulse response, i.e., $F\delta \in \ell_0(\mathbb{Z}^s)$.

**Definition 5.1.5** ($z$-transform & Convolution)**.**

1. Given a signal $c \in \ell(\mathbb{Z}^s)$, its $z$-TRANSFORM and SYMBOL are the (formal) power series

$$c^\flat(z) = \sum_{\alpha\in\mathbb{Z}^s} c(\alpha)\, z^{-\alpha}, \qquad c^\sharp(z) = \sum_{\alpha\in\mathbb{Z}^s} c(\alpha)\, z^\alpha \tag{5.1.6}$$

   respectively, defined on $\mathbb{C}_\times^s$. If $c \in \ell_0(\mathbb{Z}^s)$ then $c^\flat, c^\sharp \in \Lambda$.

2. The CONVOLUTION of two signals $c, d \in \ell(\mathbb{Z}^s)$ is defined as

$$c * d := \sum_{\alpha\in\mathbb{Z}^s} c(\cdot - \alpha) d(\alpha) = \sum_{\alpha\in\mathbb{Z}^s} d(\alpha)\, \tau^{-\alpha} c = d^\flat(\tau)\, c, \tag{5.1.7}$$

   and their CORRELATION as

$$c \star d := \sum_{\alpha\in\mathbb{Z}^s} c(\cdot + \alpha) d(\alpha) = \sum_{\alpha\in\mathbb{Z}^s} d(\alpha)\, \tau^{\alpha} c = d^\sharp(\tau)\, c. \tag{5.1.8}$$

**Remark 5.1.6.**

1. The $z$-transform of infinitely supported functions has to be taken with care as then the radius of convergence of the underlying series has to be considered.

2. Convolution and correlation are only defined if the infinite sums involved converge. This is the case if, for example $c, d \in \ell_2(\mathbb{Z}^s)$ or $d \in \ell_0(\mathbb{Z}^s)$. In the latter case convolution and correlations are the same as a difference operator.

3. Convolution is commutative, correlation is not.

4. Since

$$\begin{aligned} \|c * d\|_1 &= \sum_{\alpha\in\mathbb{Z}^s} \left| \sum_{\beta\in\mathbb{Z}^2} c(\alpha-\beta)\, d(\beta) \right| \le \sum_{\beta\in\mathbb{Z}^2} |d(\beta)| \sum_{\alpha\in\mathbb{Z}^s} |c(\alpha-\beta)| = \sum_{\beta\in\mathbb{Z}^2} |d(\beta)| \sum_{\alpha\in\mathbb{Z}^s} |c(\alpha)| \\ &= \|c\|_1 \|d\|_1, \end{aligned}$$

the correlation of two $\ell_1$ signals is $\ell_1$ again and convolution introduces a commutative MULTIPLICATION on $\ell_1(\mathbb{Z}^s)$ with neutral element $\delta$. This way, $\ell_1(\mathbb{Z}^2)$ can be turned into the so-called CONVOLUTION ALGEBRA[1].

**Exercise 5.1.3** Show that $c, d \in \ell_2(\mathbb{Z}^s)$ implies that $c * d \in \ell_1(\mathbb{Z}^s)$. ♢

The next two results are classical and extend trivially[2] from the univariate case.

**Proposition 5.1.7** (Filters & transforms)**.**

1. *For $c, d \in \ell(\mathbb{Z}^s)$ one has*

$$(c * d)^\flat = c^\flat\, d^\flat, \qquad (c \star d)^\flat = c^\flat\, d^\flat\left((\cdot)^{-1}\right) = c^\flat\, d^\sharp, \tag{5.1.9}$$

*as well as*

$$(c * d)^\sharp = c^\sharp\, d^\sharp, \qquad (c \star d)^\sharp(z) = c^\sharp\, d^\sharp\left((\cdot)^{-1}\right) = c^\sharp\, d^\flat. \tag{5.1.10}$$

2. *A linear operator $F : \ell(\mathbb{Z}) \to \ell(\mathbb{Z})$ is a filter if and only if $Fc = f * c$, $c \in \ell(\mathbb{Z}^s)$.*

**Proof:** For 1) we only prove (5.1.9) by considering

$$(c * d)^\flat(z) = \sum_{\alpha\in\mathbb{Z}^s} \sum_{\beta\in\mathbb{Z}^s} c(\alpha-\beta)\, d(\beta) z^{-\alpha} = \sum_{\beta\in\mathbb{Z}^s} d(\beta) z^{-\beta} \sum_{\alpha\in\mathbb{Z}^s} c(\alpha-\beta) z^{-\alpha+\beta} = c^\flat(z)\, d^\flat(z)$$

and, in the same way

$$\begin{aligned} (c * d)^\flat(z) &= (c * d)^\flat(z) = \sum_{\alpha\in\mathbb{Z}^s} \sum_{\beta\in\mathbb{Z}^s} c(\alpha+\beta)\, d(\beta) z^{-\alpha} = \sum_{\beta\in\mathbb{Z}^s} d(\beta) z^{\beta} \sum_{\alpha\in\mathbb{Z}^s} c(\alpha+\beta) z^{-\alpha-\beta} \\ &= c^\flat(z)\, d^\flat(z^{-1}). \end{aligned}$$

For 2), "⇒", we use the trivial reformulation

$$c = \sum_{\alpha\in\mathbb{Z}^s} c(\alpha)\tau^{-\alpha}\delta$$

and the commutativity property (5.1.5) to find that

$$Fc = F\left(\sum_{\alpha\in\mathbb{Z}^s} c(\alpha)\tau^{-\alpha}\delta\right) = \sum_{\alpha\in\mathbb{Z}^s} c(\alpha) F\tau^{-\alpha}\delta = \sum_{\alpha\in\mathbb{Z}^s} c(\alpha)\tau^{-\alpha} F\delta = \sum_{\alpha\in\mathbb{Z}^s} c(\alpha)\tau^{-\alpha} f = f * c$$

[1] An ALGEBRA is a vector space, here even a normed one, with compatible multiplication.
[2] Only by formal extension.

while the converse, "⇐", follows from the fact that

$$\tau^\alpha (f * c) = \left(\tau^\alpha f^\flat(\tau)\right) c = \left(f^\flat(\tau)\, \tau^\alpha\right) = f^\flat(\tau)\left(\tau^\alpha c\right) = f * (\tau^\alpha c),$$

due to the commutativity of polynomial multiplication. □

**Exercise 5.1.4** Prove (5.1.10) without being surprised that the proof is extremely similar to that of (5.1.9). ♢

In the sequel we will be interested in FIR filters only, so we record the following consequence of Proposition 5.1.7 for further convenience.

**Corollary 5.1.8.** *Any FIR filter $F$ can be written as $Fc = f * c$, $f \in \ell_0(\mathbb{Z}^s)$ or as*

$$Fc = f^\flat(\tau)\, c, \qquad c \in \ell(\mathbb{Z}^s). \tag{5.1.11}$$

**Proof:** By (5.1.7) we have

$$f * c = c * f = f^\flat(\tau)\, c,$$

(5.1.11) follows immediately. □

## 5.2 Difference equations and their homogeneous solutions

A difference equation is an expression of the form

$$q(\tau)\, u = v, \qquad q \in \Pi, \qquad u, v \in \ell(\mathbb{Z}^s). \tag{5.2.1}$$

More precisely, it is a *linear* difference equation with CONSTANT coefficients but since we neither consider nonlinear equations nor variable coefficients[3], we will keep the above shorter definition.

As usual with linear operators, we are interested in the KERNEL of the operator, i.e., the solutions of the homogeneous equation

$$q(\tau)\, u = 0.$$

Some of those can be determined quite easily.

**Definition 5.2.1.** For $\theta \in \mathbb{C}^s_\times$, we define the signal $c_\theta \in \ell(\mathbb{Z}^s)$ as

$$c_\theta = \theta^{(\cdot)} = \left(\alpha \mapsto \theta^\alpha : \alpha \in \mathbb{Z}^s\right). \tag{5.2.2}$$

We call $c_\theta$ an EXPONENTIAL SIGNAL.

**Remark 5.2.2** (Exponential signals)**.**

1. We require $\theta_j \neq 0$ since then we do not have the problem to explain what $0^0$ means; moreover, we will soon see that $\theta$ will correspond to a (forbidden) zero of a Laurent polynomial.
2. With $\omega := \log\theta := \left(\log\theta_j : j = 1, \ldots, s\right)$ we can also write $c_\theta = e_\omega = e^{\omega^T(\cdot)}$ and have yet another reason to request $\theta \in \mathbb{C}^s_\times$.

[3] Anyway, this is more common in the context of partial differential equations.

Now we make a simple and elementary computation that will turn out to be fundamental for all that follows in this section. For $\theta \in \mathbb{C}_\times^s$ we consider

$$q(\tau)c_\theta = \sum_{\alpha\in\mathbb{Z}^s} q_\alpha \tau^\alpha \theta^{(\cdot)} = \sum_{\alpha\in\mathbb{Z}^s} q_\alpha \theta^{(\cdot)+\alpha} = \theta^{(\cdot)} \sum_{\alpha\in\mathbb{Z}^s} q_\alpha \theta^\alpha = q(\theta)\, c_\theta,$$

which we can summarize as follows.

**Proposition 5.2.3** (Difference operators & exponential signals)**.** *The exponential signal $c_\theta$ is an eigenvector of any difference operator $c \mapsto q(\tau)c$, $q \in \Pi$, with eigenvalue $q(\theta)$. In particular,*

$$q(\tau)c_\theta = 0 \qquad \Leftrightarrow \qquad q(\theta) = 0. \tag{5.2.3}$$

Despite its simple proof, Proposition 5.2.3 is of fundamental importance for what follows.

**Remark 5.2.4.**

1. The kernel of the difference operator is obviously related to the zeros of the polynomial $q$.
2. Since, by (5.1.11), any FIR filter is equivalent to a difference equation defined by the $z$-transform of its impulse response, the two concepts are mainly equivalent.
3. Any difference operator or filter that is not not the identity, i.e., $q = 1$ or $f = \delta$, has a $z$-transform that vanishes on a whole variety and therefore the kernel is an infinite dimensional subspace. This is *not* the case for $s = 1$.
4. Indeed, the case $s = 1$ is very classic and investigated for quite some time, cf.[Goldberg, 1958, Jordan, 1965]

### 5.2.1 Systems of difference equations

Obviously, a single difference equation will never have a finite dimensional kernel as single polynomial can have no simple zeros. To that end, the multivariate case requires us to consider several equations *simultaneously*.

**Definition 5.2.5.** A SYSTEM OF DIFFERENCE EQUATIONS is given as a *finite* set $Q \subset \Pi$ of polynomials and $v \in \ell(\mathbb{Z}^s)^Q$ and consists of finding $u \in \ell(\mathbb{Z}^s)$ such that

$$Q(\tau)u = v, \qquad \text{i.e.,} \qquad q(\tau)u = v_q, \quad q \in Q. \tag{5.2.4}$$

A system of HOMOGENEOUS DIFFERENCE EQUATIONS is the case when $v = 0$.

**Remark 5.2.6.** Homogeneous difference equations depend on $\langle Q\rangle$, not on the specific choice $Q$. Indeed, since trivially $p(\tau)0 = 0$, any solution of $Q(\tau)u = 0$ satisfies

$$0 = g_q(\tau)0 = g_q(\tau)q(\tau)0 = \big(g_q q\big)(\tau) = 0 \qquad \Rightarrow \qquad \left(\sum_{q\in Q} g_q g\right)(\tau)u = 0, \qquad g_q \in \Pi,\ q \in Q,$$

that is,

$$Q(\tau)u = 0 \qquad \Leftrightarrow \qquad \langle Q\rangle(\tau)u = 0. \tag{5.2.5}$$

We will be interested in homogeneous difference equations whose solution space is *finite dimensional*. To that end, we use a straightforward generalization of Proposition 5.2.3.

**Proposition 5.2.7.** *For any finite $Q \subset \Pi$ and $\theta \in \mathbb{C}_\times^s$ we have that*

$$Q(\tau)c_\theta = 0 \qquad \Leftrightarrow \qquad \theta \in Z(\langle Q \rangle). \tag{5.2.6}$$

**Proof:** From Proposition 5.2.3 we know that $q(\tau)c_\theta = 0$ iff $q(\theta) = 0$, hence

$$Q(\tau)c_\theta = 0 \qquad \Leftrightarrow \qquad q(\tau)c_\theta = 0, \quad q \in Q, \qquad \Leftrightarrow \qquad Q(\theta) = 0 \qquad \Leftrightarrow \qquad \langle Q \rangle(\theta) = 0,$$

which gives (5.2.6). □

This already allows us to characterize kernels of partial difference equations in a simple[4] situation.

**Definition 5.2.8.** A subspace $\mathscr{F} \subset \ell(\mathbb{Z}^s)$ is called SHIFT INVARIANT if $\tau^\alpha \mathscr{F} \subseteq \mathscr{F}$, $\alpha \in \mathbb{Z}^s$. Given $F \subset \ell(\mathbb{Z}^s)$, the shift invariant space generated by $F$ is

$$S(F) := \operatorname{span}\left\{\tau^\alpha f : \alpha \in \mathbb{Z}^s,\, f \in F\right\}.$$

If $\#F = 1$, i.e., $F = \{f\}$, then the resulting shift invariant space $S(f)$ spanned by a single signal is sometimes also called a PRINCIPAL SHIFT INVARIANT SPACE[5].

**Example 5.2.9.** The simplest example of a principal shift invariant space is again the signal $c_\theta$ from (5.2.2). Since

$$\tau^\alpha c_\theta = \theta^{(\cdot)+\alpha} = \theta^\alpha c_\theta,$$

is follows that $\dim S(c_\theta) = 1$ for any $\theta \in \mathbb{C}_\times^s$. But also the converse is true. If $\dim S(f) = 1$ for some $f$, then there exist $\theta_j \in \mathbb{C}$ such that

$$\theta_j f = \tau^{\epsilon_j} f = f\left(\cdot + \epsilon_j\right),$$

yielding $\tau^\alpha f = \theta^\alpha f$, $\alpha \in \mathbb{Z}^s$, and thus, evaluating at 0,

$$f(\alpha) = \left(\tau^\alpha f\right)(0) = \theta^\alpha f(0)$$

from which is follows that $f = \theta^{(\cdot)}$ and that $\theta \in \mathbb{C}_\times^s$ as otherwise $f = 0$ and $\dim S(f) = 0$.

**Corollary 5.2.10.** *For $f \in \ell(\mathbb{Z}^s)$ one has that*

$$\dim S(f) = 1 \qquad \Leftrightarrow \qquad f = c_\theta, \quad \theta \in \mathbb{C}_\times^s. \tag{5.2.7}$$

But there is also a connection between solutions of systems of homogeneous difference equations and shift invariance that we will explore next. Let us begin with a simple observation.

**Lemma 5.2.11.** *The space* $\ker Q(\tau)$ *is shift invariant.*

**Proof:** For $f \in \ker Q(\tau) = \ker \langle Q \rangle(\tau)$ and $\alpha \in \mathbb{N}_0^s$ we have that

$$Q(\tau)\left(\tau^\alpha f\right) = \left((\cdot)^\alpha Q\right)(\tau) f = 0$$

since $(\cdot)^\alpha Q \subset \langle Q \rangle$ and due to (5.2.5). □

**Lemma 5.2.12.** *If $0 \neq f \in \ell_0(\mathbb{Z}^s)$ then* $\dim S(f) = \infty$.

**Proof:** Since $\Omega := \operatorname{supp} f \subset \mathbb{Z}^s$ is a finite set, there exists $\alpha \in \mathbb{Z}^s$ such that $\tau^\alpha \Omega \cap \Omega = \emptyset$, hence all the sets $\tau^{k\alpha}\Omega$, $k \in \mathbb{Z}$, are disjoint and $\tau^{k\alpha} f$ linearly independent. □

[4] In a double sense: it is simple since the zeros are simple.

[5] In obvious analogy to the concept of a principal ideal.

### 5.2.2 Stirling numbers and Stirling operators

We continue defining some numbers that will be of particular use in what follows.

**Definition 5.2.13.** The multivariate STIRLING NUMBERS of the second kind KARAMATA'S NOTATION, cf. [Graham et al., 1998, p. 257ff] are defined as *differences of zero*

$$\left\{ \begin{matrix} \nu \\ \kappa \end{matrix} \right\} := \frac{1}{\kappa!} \Delta^\kappa 0^\nu := \frac{1}{\kappa!} \left( \Delta^\kappa (\cdot)^\nu \right)(0), \qquad \kappa, \nu \in \mathbb{N}_0^s. \tag{5.2.8}$$

The Stirling numbers of the first kind are defined as

$$\left[ \begin{matrix} \nu \\ \kappa \end{matrix} \right] := \frac{1}{\kappa!} \left( D^\kappa (\cdot)_\nu \right)(0), \tag{5.2.9}$$

with the FALLING FACTORIALS or POCHHAMMER SYMBOLS[6]

$$(\cdot)_\alpha = \prod_{j=1}^{s} \prod_{k=0}^{\alpha_j - 1} \left( (\cdot)_j - k \right), \qquad \alpha \in \mathbb{N}_0^s \tag{5.2.10}$$

and $D^\alpha = \frac{\partial^{|\alpha|}}{\partial x^\alpha}$ as abbreviation for the partial derivatives.

Obviously we have that

$$0 = \left\{ \begin{matrix} \nu \\ \kappa \end{matrix} \right\} = \left[ \begin{matrix} \nu \\ \kappa \end{matrix} \right], \qquad \nu \not\geq \kappa, \tag{5.2.11}$$

which allows us to extend Sterling numbers to arbitrary pairs $\nu, \kappa$ by convention, where nonzero values only occur for $\kappa \leq \nu$.

**Remark 5.2.14.** Stirling numbers are a well investigated topic since the play a quite important role in analysis and analytic number theory. Therefore I refer to a statement from [Gould, 1971]: *"...aber es mag von Interesse sein, daß mindestestens tausend Abhandlungen in der Literatur existieren, die sich mit den Stirlingschen Zahlen beschäftigen. Es ist also sehr schwer, etwas Neues über die Stirlingschen Zahlen zu entdecken."*

Note that by (5.1.4)

$$\Delta^\kappa (\cdot)^\nu = (\tau - I)^\kappa (\cdot)^\nu = (-1)^{|\kappa|} \sum_{\alpha \leq \kappa} \binom{\kappa}{\alpha} (-1)^{|\alpha|} (\cdot + \alpha)^\nu = (-1)^{|\kappa|} \sum_{\alpha \leq \kappa} \binom{\kappa}{\alpha} (-1)^{|\alpha|} \sum_{\beta \leq \nu} \binom{\nu}{\beta} \alpha^{\nu - \beta} (\cdot)^\beta$$

and evaluation at 0 leaves only the term $\beta = 0$ in the second sum and gives the explicit formula

$$\left\{ \begin{matrix} \nu \\ \kappa \end{matrix} \right\} = \sum_{\alpha \leq \kappa} \binom{\kappa}{\alpha} (-1)^{|\kappa| - |\alpha|} \alpha^\nu. \tag{5.2.12}$$

Moreover, the Leibniz rule for the difference,

$$\Delta^\alpha (fg) = \sum_{\beta \leq \alpha} \Delta^\beta f \, \tau^\beta \Delta^{\alpha - \beta} g, \qquad \alpha \in \mathbb{N}_0^s, \tag{5.2.13}$$

yields that

$$\Delta^\kappa (\cdot)^{\nu + \epsilon_j} = \Delta^\kappa \left( (\cdot)^\nu (\cdot)^{\epsilon_j} \right) = \Delta^\kappa (\cdot)^\nu \underbrace{\tau^\kappa (\cdot)^{\epsilon_j}}_{= \kappa_j} + \Delta^{\kappa - \epsilon_j} (\cdot)^\nu \tau^{\kappa - \epsilon_j} \underbrace{\Delta^{\epsilon_j} (\cdot)^{\epsilon_j}}_{=1} = \kappa_j \Delta^\kappa (\cdot)^\nu + \Delta^{\kappa - \epsilon_j} (\cdot)^\nu$$

[6] The integer version of (4.3.24).

and thus the recurrence relation

$$\left\{ \begin{matrix} \nu+\epsilon_j \\ \kappa \end{matrix} \right\} = \left(\kappa_j \Delta^{\kappa} (\cdot)^{\nu} + \Delta^{\kappa-\epsilon_j} (\cdot)^{\nu}\right)(0) = \kappa_j \left\{ \begin{matrix} \nu \\ \kappa \end{matrix} \right\} + \left\{ \begin{matrix} \nu \\ \kappa-\epsilon_j \end{matrix} \right\}. \tag{5.2.14}$$

**Exercise 5.2.1** Prove (5.2.13) by using the univariate formula

$$\Delta^n (fg) = \sum_{k=0}^{n} \left(\Delta^k f\right)\left(\tau^k \Delta^{n-k} g\right),$$

cf. [Boor, 2005]. ♢

Next recall the Taylor formula and the Newton formula of interpolation at integers giving

$$f = \sum_{\alpha\in\mathbb{N}_0^s} \frac{1}{\alpha!} D^{\alpha} f(0)\, (\cdot)^{\alpha} = \sum_{\alpha\in\mathbb{N}_0^s} \frac{1}{\alpha!} \Delta^{\alpha} f(0)\, (\cdot)_{\alpha}, \qquad f\in\Pi,$$

and switch the roles of functionals and polynomials between these two.

**Definition 5.2.15.** The STIRLING OPERATOR of the first kind $L_1 : \Pi \to \Pi$ and the one of the second kind, $L_2 : \Pi \to \Pi$, are defined as

$$L_1 f := \sum_{\alpha\in\mathbb{N}_0^s} \frac{1}{\alpha!} D^{\alpha} f(0)\, (\cdot)_{\alpha}, \qquad L_2 f := \sum_{\alpha\in\mathbb{N}_0^s} \frac{1}{\alpha!} \Delta^{\alpha} f(0)\, (\cdot)^{\alpha}. \tag{5.2.15}$$

The name is due to the following observations.

**Proposition 5.2.16.** *The operators $L_1, L_2$ satisfy*

$$L_1 = L_2^{-1} \tag{5.2.16}$$

*and*

$$\left(L_1 f\right)'_{\alpha} = \sum_{\beta\in\mathbb{N}_0^s} \left[ \begin{matrix} \beta \\ \alpha \end{matrix} \right] f'_{\alpha}, \qquad \left(L_2 f\right)_{\alpha} = \sum_{\beta\in\mathbb{N}_0^s} \left\{ \begin{matrix} \beta \\ \alpha \end{matrix} \right\} f_{\alpha}, \qquad f = \sum_{\alpha\in\mathbb{N}_0^s} f_{\alpha} (\cdot)^{\alpha} = \sum_{\alpha\in\mathbb{N}_0^s} f'_{\alpha} (\cdot)_{\alpha}. \tag{5.2.17}$$

**Proof:** For $f\in\Pi$ we have that

$$L_1 L_2 f = \sum_{\alpha,\beta\in\mathbb{N}_0^s} \frac{1}{\alpha!} \Delta^{\alpha} f(0) \frac{1}{\beta!} \underbrace{\left(D^{\beta} (\cdot)^{\alpha}\right)(0)}_{=\alpha!\delta_{\alpha,\beta}} (\cdot)_{\beta} = \sum_{\alpha\in\mathbb{N}_0^s} \frac{1}{\alpha!} \Delta^{\alpha} f(0)\, (\cdot)_{\alpha} = f,$$

which proves (5.2.16). For the explicit expression of the second kind operator we just note that the definition yields

$$L_2 f = \sum_{\alpha,\beta\in\mathbb{N}_0^s} f_{\alpha} \frac{1}{\beta!} \left(\Delta^{\beta} (\cdot)^{\alpha}\right)(0)\, (\cdot)^{\beta} = \sum_{\beta\in\mathbb{N}_0^s} (\cdot)^{\beta} \sum_{\alpha\in\mathbb{N}_0^s} \left\{ \begin{matrix} \beta \\ \alpha \end{matrix} \right\} f_{\alpha}.$$

The expression for $L_1 f$ is done in the same way. □

**Corollary 5.2.17.** *We have*

$$\sum_{\gamma\in\mathbb{N}_0^s} \left\{ \begin{matrix} \alpha \\ \gamma \end{matrix} \right\} \left[ \begin{matrix} \gamma \\ \beta \end{matrix} \right] = \delta_{\alpha,\beta}, \qquad \alpha,\beta\in\mathbb{N}_0^s. \tag{5.2.18}$$

**Definition 5.2.18.** In the sequel we use $L := L_2$ with $L^{-1} = L_1$.

Finally an important property for multiplicity spaces: the Stirling operator respects $D$ invariance.

**Proposition 5.2.19.** *$\mathscr{Q} \subset \Pi$ is $D$-invariant if and only if $L\mathscr{Q}$ is $D$-invariant.*

**Proof:** For "$\Rightarrow$" we choose $f \in \Pi$ and $\alpha \in \mathbb{N}_0^s$ and compute

$$\begin{aligned} D^\alpha Lf &= \sum_{\beta\in\mathbb{N}_0^s} \frac{1}{\beta!}\Delta^\beta f(0)\, D^\alpha (\cdot)^\beta = \sum_{\beta\geq\alpha} \frac{1}{\beta!}\Delta^\beta f(0) \frac{\beta!}{(\beta-\alpha)!}(\cdot)^{\beta-\alpha} \\ &= \sum_{\beta\geq\alpha} \Delta^{\beta-\alpha}\Delta^\alpha f(0)\frac{1}{(\beta-\alpha)!}(\cdot)^{\beta-\alpha} = \sum_{\beta\in\mathbb{N}_0^s} \frac{1}{\beta!}\Delta^\beta\left(\Delta^\alpha f\right)(0)\,(\cdot)^\beta = L\Delta^\alpha f, \end{aligned}$$

i.e.,

$$D^\alpha L = L\Delta^\alpha, \qquad \alpha \in \mathbb{N}_0^s, \tag{5.2.19}$$

From (5.2.19) we also get that

$$L^{-1}D^\alpha = L^{-1}D^\alpha LL^{-1} = L^{-1}L\Delta^\alpha L^{-1} = \Delta^\alpha L^{-1}, \qquad \alpha \in \mathbb{N}_0^s, \tag{5.2.20}$$

and (5.2.19) implies that $D^\alpha Lq = L\Delta^\alpha q \in L\mathscr{Q}$, $q \in \mathscr{Q}$. whenever $\mathscr{Q}$ is $D$-invariant, hence shift invariant, see Proposition 3.2.3 since then $\Delta^\alpha q \in \mathscr{Q}$.

Conversely, "$\Leftarrow$" follows since for any $q' = Lq \in L\mathscr{Q}$, $q \in \mathscr{Q}$, that

$$\Delta^\alpha q = \Delta^\alpha L^{-1}Lq = \Delta^\alpha L^{-1}q' = L^{-1}\underbrace{D^\alpha q'}_{\in L\mathscr{Q}} \in L^{-1}L\mathscr{Q} = \mathscr{Q}$$

which completes the proof. □

The last concept that we introduce in this section is a variation of the derivative operator that will turn out to be quite useful in simplifying things.

**Definition 5.2.20.** The modified partial differential operator is defined as

$$\frac{\partial_*}{\partial_* x_j} := (\cdot)_j \frac{\partial}{\partial x_j}, \quad j = 1,\dots,s, \qquad \text{and} \qquad D_*^\alpha := \frac{\partial_*^{|\alpha|}}{\partial_* x^\alpha}, \quad \alpha \in \mathbb{N}_0^s. \tag{5.2.21}$$

This operator is sometimes also called the $\theta$-OPERATOR.

**Lemma 5.2.21.** *The $\theta$-operator is well defined, in particular*

$$D_*^\alpha = \sum_{\beta\leq\alpha} \left\{ \begin{matrix} \alpha \\ \beta \end{matrix} \right\} (\cdot)^\beta D^\beta, \qquad \alpha \in \mathbb{N}_0^s. \tag{5.2.22}$$

**Proof:** We use induction on $|\alpha|$, where the case $|\alpha| = 1$ is just the definition in (5.2.21). Moreover, we note that for $j = 1,\dots,s$, taking into account that $\left\{ \begin{matrix} \alpha \\ \beta \end{matrix} \right\} \neq 0$ only for $\beta \leq \alpha$, cf. (5.2.8),

$$\begin{aligned} D_*^{\alpha+\epsilon_j} &= (\cdot)_j \frac{\partial}{\partial x_j} D_\alpha^* = (\cdot)_j \frac{\partial}{\partial x_j} \sum_{\beta\leq\alpha} \left\{ \begin{matrix} \alpha \\ \beta \end{matrix} \right\} (\cdot)^\beta D^\beta = \sum_{\beta\leq\alpha} \left\{ \begin{matrix} \alpha \\ \beta \end{matrix} \right\} \left( \beta_j (\cdot)^\beta D^\beta + (\cdot)^{\beta+\epsilon_j} D^{\beta+\epsilon_j} \right) \\ &= \sum_{\beta\leq\alpha} \beta_j \left\{ \begin{matrix} \alpha \\ \beta \end{matrix} \right\} (\cdot)^\beta D^\beta + \sum_{\beta\leq\alpha+\epsilon_j} \left\{ \begin{matrix} \alpha \\ \beta-\epsilon_j \end{matrix} \right\} (\cdot)^\beta D^\beta \\ &= \sum_{\beta\leq\alpha+\epsilon_j} \left( \beta_j \left\{ \begin{matrix} \alpha \\ \beta \end{matrix} \right\} + \left\{ \begin{matrix} \alpha \\ \beta-\epsilon_j \end{matrix} \right\} \right) (\cdot)^\beta D^\beta = \sum_{\beta\leq\alpha+\epsilon_j} \left\{ \begin{matrix} \alpha+\epsilon_j \\ \beta \end{matrix} \right\} (\cdot)^\beta D^\beta, \end{aligned}$$

which advances the induction hypothesis and completes the proof. □

**Corollary 5.2.22.** *For $q \in \Pi$ and $x \in \mathbb{C}_\times^s$ we have that*

$$q(D_*)f(x) = \big(Lq(xD)\big)f(x), \qquad f \in \Pi. \tag{5.2.23}$$

**Proof:** For (5.2.23) we get

$$\begin{aligned} q(D_*) &= \sum_{\alpha\in\mathbb{N}_0^s} q_\alpha D_*^\alpha = \sum_{\alpha\in\mathbb{N}_0^s} q_\alpha \sum_{\beta\le\alpha} \left\{ \begin{matrix} \alpha \\ \beta \end{matrix} \right\} (\cdot)^\beta D^\beta = \sum_{\beta\in\mathbb{N}_0^s} \left( \sum_{\alpha\ge\beta} \left\{ \begin{matrix} \alpha \\ \beta \end{matrix} \right\} q_\alpha \right) (\cdot)^\beta D^\beta \\ &= \sum_{\beta\in\mathbb{N}_0^s} \big(Lq\big)_\beta (\cdot)^\beta D^\beta = Lq(xD) \end{aligned}$$

by straightforward computation. □

### 5.2.3 Exponential polynomials and multiplicities

**Definition 5.2.23.** An EXPONENTIAL POLYNOMIAL is a function of the form

$$x \mapsto p(x)e_y(x) = p(x)\, e^{y^T x}, \qquad p \in \Pi, \quad y \in \mathbb{C}^s. \tag{5.2.24}$$

The restriction of an exponential polynomial to $\mathbb{Z}^s$ is an exponential polynomial signal.

The next result gives a fundamental relationship between the exponential polynomial signals in the kernel of a difference operator and the zeros of the associated polynomial.

**Theorem 5.2.24.** *Let $\Theta \subset \mathbb{C}_\times^s$, $\#\Theta < \infty$, and $\mathscr{Q}_\theta \subset \Pi$ be finite dimensional $D$-invariant spaces. Then, for any $f \in \Pi$,*

$$f(\tau)(\mathscr{Q}_\theta c_\theta) = 0, \quad \theta \in \Theta, \qquad \Leftrightarrow \qquad \big(L\mathscr{Q}_\theta(\theta D)f\big)(\theta) = 0, \quad \theta \in \Theta. \tag{5.2.25}$$

**Proof:** For $f, q \in \Pi$ and $\theta \in \mathbb{C}^s$ we consider

$$\begin{aligned} f(\tau)\big(p\, c_\theta\big) &= \sum_{\alpha\in\mathbb{N}_0^s} f_\alpha \tau^\alpha \big(q\theta^{(\cdot)}\big) = \sum_{\alpha\in\mathbb{N}_0^s} f_\alpha p(\cdot+\alpha)\theta^{(\cdot+\alpha)} \\ &= \theta^{(\cdot)} \sum_{\alpha\in\mathbb{N}_0^s} f_\alpha \sum_{\beta\in\mathbb{N}_0^s} \frac{1}{\beta!} \Delta^\beta(\tau^{(\cdot)}q)(0)\,(\alpha)_\beta \theta^\alpha \\ &= \theta^{(\cdot)} \sum_{\alpha\in\mathbb{N}_0^s} f_\alpha \sum_{\beta\in\mathbb{N}_0^s} \frac{1}{\beta!} \Delta^\beta(\tau^{(\cdot)}q)(0)\theta^\beta \Big(D^\beta(\cdot)^\alpha\Big)(\theta) \\ &= \theta^{(\cdot)} \sum_{\beta\in\mathbb{N}_0^s} \frac{1}{\beta!} \Delta^\beta(\tau^{(\cdot)}q)(0)\,(\theta D)^\beta \left( \sum_{\alpha\in\mathbb{N}_0^s} f_\alpha (\cdot)^\alpha \right)(\theta) \\ &= \theta^{(\cdot)} \big(L\tau^{(\cdot)}q\big)(\theta D)\, f(\theta), \end{aligned}$$

i.e.,

$$f(\tau)\big(q\, c_\theta\big)(\alpha) = \theta^\alpha \big(L\tau^\alpha q\big)(\theta D)\, f(\theta), \qquad \alpha \in \mathbb{Z}^s. \tag{5.2.26}$$

Since $\tau^\alpha q \in \mathscr{Q}_\theta$ for $q \in \mathscr{Q}_\theta$ and $\alpha \in \mathbb{Z}^s$, the direction "$\Leftarrow$" follows directly from (5.2.26). Conversely, to obtain "$\Rightarrow$", we simply set $\alpha = 0$ in (5.2.26) and find that

$$0 = f(\tau)(qc_\theta)(0) = (Lq)(\theta D)f(\theta), \qquad \theta \in \Theta,$$

which is the right hand side of (5.2.25). □

The $\theta$-operator now allows us to formulate Theorem 5.2.24 in a shorter and more elegant way.

**Corollary 5.2.25.** *Let $\Theta \subset \mathbb{C}_\times^s$, $\#\Theta < \infty$, and $\mathscr{Q}_\theta \subset \Pi$ be finite dimensional $D$-invariant spaces. Then, for any $f \in \Pi$,*

$$f(\tau)(\mathscr{Q}_\theta c_\theta) = 0, \quad \theta \in \Theta, \qquad \Leftrightarrow \qquad \mathscr{Q}_\theta(D_*) f(\theta) = 0, \quad \theta \in \Theta. \tag{5.2.27}$$

Moreover, this can be extended to systems of difference equations to identify at least some of the homogeneous solutions.

**Definition 5.2.26.** For a finite dimensional $D$-invariant subspace $\mathscr{Q} \subset \Pi$ and $\theta \in \mathbb{C}_\times^s$ define the $\theta$ MULTIPLICITY SPACE

$$\mathscr{Q}^* := L^{-1}\mathscr{Q}\left(\theta^{-1}\cdot\right). \tag{5.2.28}$$

so that

$$\mathscr{Q}(D) f(\theta) = \mathscr{Q}^*(D_*) f(\theta).$$

**Corollary 5.2.27.** *Let $F \subset \Pi$ be such that $\langle F \rangle$ is a zero dimensional ideal. Then*

$$\ker F(\tau) \supseteq \operatorname{span}\left\{\mathscr{Q}_\theta^* e_\theta : \theta \in Z(\langle F \rangle)\right\}, \tag{5.2.29}$$

*where $\mathscr{Q}_\theta^*$ stands for the $\theta$ multiplicity space of the zero $\theta \in \mathbb{C}_\times^s$.*

**Remark 5.2.28.** If $F$ has common zeros with a zero component, then these zeros do not contribute to the solution space for the difference equation as the respective exponential sequences are not well-defined.

### 5.2.4 Finite dimensional shift invariant spaces

In Section 5.2.3, especially in Corollary 5.2.27, we saw that all exponential polynomial sequences corresponding to the zeros of the ideal and their multiplicities are homogeneous solutions of the difference equation. We now head for the converse, considering more carefully finite dimensional spaces of homogeneous solutions of difference equations.

We start with some simple observations.

**Lemma 5.2.29.**

1. *For any $F \subset \Pi$ the space $\ker F(\tau) \subseteq \ell(\mathbb{Z}^s)$ is shift invariant.*
2. *If $\mathscr{U} \subset \ell(\mathbb{Z}^s)$ is a shift invariant subspace, then*

$$\mathscr{I}(\mathscr{U}) := \left\{f : f(\tau)\mathscr{U} = 0\right\} \tag{5.2.30}$$

*is a Laurent ideal.*

**Proof:** By (5.2.5), $\ker F(\tau) = \ker \langle F \rangle(\tau)$ and any $u \in \ker \langle F \rangle(\tau)$ satisfies

$$0 = \left((\cdot)^\alpha F\right)(\tau) u = F(\tau)\left(\tau^\alpha u\right),$$

hence $\tau^\alpha u \in \ker F(\tau)$ which is 1).

For 2) we note that the shift invariance of $\mathscr{U}$ with respect to *all*[7] multiinteger shifts implies that $\tau^\alpha \mathscr{U} \subset \mathscr{U}$, $\alpha \in \mathbb{Z}^s$, hence $q(\tau)\mathscr{U} \subset \mathscr{U}$ for $q \in \Lambda$, and any $f$ such that $f(\tau)\mathscr{U} = 0$ satisfies

$$0 = f(\tau) q_f(\tau)\mathscr{U} = \left(q_f f\right)(\tau)\mathscr{U}, \quad q_f \in \Pi \qquad \Rightarrow \qquad \left(\sum_{f \in F} q_f f\right)(\tau)\mathscr{U}, \quad q_f \in \Lambda, f \in F,$$

[7] And not only those with $\alpha \in \mathbb{N}_0^s$

so that $\mathscr{I}(\mathscr{U})$ is indeed a Laurent ideal. □

Recall from Section 2.1.6 that in order to determine a basis for a Laurent ideal, we first compute the polynomial part $P(\mathscr{I}(\mathscr{U})) = \mathscr{I}(\mathscr{U}) \cap \Pi$ of the Laurent ideal and then a basis of that polynomial ideal.

**Definition 5.2.30.** For a Laurent ideal $\mathscr{I} \subset \Lambda$ we define the ZERO SET

$$Z(\mathscr{I}) = Z(P(\mathscr{I})) \tag{5.2.31}$$

and the QUOTIENT SPACE

$$\Pi/\mathscr{I} := \Pi/P(\mathscr{I}). \tag{5.2.32}$$

**Remark 5.2.31** (Laurent ideals)**.**

1. Recall that Remark 2.1.36 describes a *constructive* way to compute $P(\mathscr{I})$ for a given Laurent ideal by saturating quotient ideals, see also [Möller and Sauer, 2004].
2. This way, we can even compute a $\Gamma$-basis for a Laurent ideal: first we saturate the ideal basis into a basis of $P(\mathscr{I})$ and then we compute the $\Gamma$-basis from this saturated basis.
3. Statement 1) of Proposition 2.1.38 also tells us that $Z(\mathscr{I}) \subset \mathbb{C}^s_\times$, that is, all common zeros of the Laurent ideal according to (5.2.31) have only nonzero components. There are no *spurious zeros*.

**Lemma 5.2.32.** *If $\mathscr{U}$ is finite dimensional, then* $\dim \Pi/\mathscr{I}(\mathscr{U}) = \dim \mathscr{U}$ *and* $\mathscr{U} = \ker F(\tau)$ *for any basis $F$ of $\mathscr{I}(\mathscr{U})$.*

**Proof:** Let $U \subset \ell(\mathbb{Z}^s)$ be a basis of $\mathscr{U}$ and $P$ a basis of $\Pi/\mathscr{I}(\mathscr{U})$ and consider the matrix

$$P(\tau)U = \left( p(\tau)u : \begin{array}{c} p \in P \\ u \in U \end{array} \right). \tag{5.2.33}$$

The rank of this matrix is at most $\dim \mathscr{U}$ and if $\#P = \dim \Pi/\mathscr{I}(\mathscr{U})$ were $> \dim \mathscr{U}$, then there would exist $0 \neq y \in \mathbb{C}^P$ such that

$$0 = y^T P(\tau)U = \left( \sum_{p \in P} y_p\, p \right)(\tau)u \qquad \Rightarrow \qquad y \cdot P \in \mathscr{I}(\mathscr{U}) \qquad \Rightarrow \qquad y = 0,$$

which would be a contradiction. Hence $\#P \leq \#U$. If, on the other hand, $\#P < \#U$, then there exists $0 \neq y \in \mathbb{C}^U$ such that

$$0 = P(\tau)(y \cdot U) = (\mathscr{I}(\mathscr{U}))(\tau)(y \cdot U) \qquad \Rightarrow \qquad f(\tau)(y \cdot U) = 0, \quad f \in \Pi,$$

and choosing $f = (\cdot)^\alpha$ yields $(y \cdot U)(\alpha) = 0$, $\alpha \in \mathbb{Z}^s$ yields $y \cdot U = 0$ which is a contradiction. Hence, $\dim \Pi/\mathscr{I}(\mathscr{U}) = \dim \mathscr{U}$. Moreover, the definition in (5.2.30) yields that $\mathscr{U} \subseteq \ker F(\tau)$ for any $F \subset \mathscr{I}(\mathscr{U})$, in particular for a basis of $\mathscr{I}(\mathscr{U})$. But since

$$\dim \Pi/\langle F \rangle = \dim \Pi/\mathscr{I}(\mathscr{U}) = \dim \mathscr{U},$$

the two spaces have to coincide, yielding $\mathscr{U} = \ker F(\tau)$. □

This Lemma enables us to characterize the homogeneous solutions of partial difference equations.

**Theorem 5.2.33.** *If the system of difference equations $F(\tau)u = 0$ has a finite dimensional space* $\ker F(\tau) =: \mathscr{U} \subset \ell(\mathbb{Z}^s)$ *of homogeneous solution, then*

$$\mathscr{U} = \sum_{\theta \in Z(\langle F \rangle)} \mathscr{Q}_\theta^* e_\theta, \qquad \langle F \rangle = \bigcap_{\theta \in Z(\langle F \rangle)} \ker \delta_\theta \circ \mathscr{Q}_\theta^*(D_*). \tag{5.2.34}$$

**Proof:** Corollary 5.2.27 already states that $\mathscr{U} \subseteq \sum_\theta \mathscr{Q}_\theta^* e_\theta$. But Lemma 5.2.32 says that

$$\dim \mathscr{U} = \dim \Pi / \langle F \rangle = \sum_{\theta \in Z(\langle F \rangle)} \dim \mathscr{Q}_\theta,$$

hence both spaces have the same dimension and therefore must agree. □

**Remark 5.2.34.** The operators $\mathscr{Q}_\theta^*$ are well-defined since the polynomial part of the Laurent ideal ensures that $\theta \in \mathbb{C}_\times^s$.

**Theorem 5.2.35.** *Let $\mathscr{U} \subset \ell(\mathbb{Z}^s)$ be a finite dimensional shift invariant space. Then $\mathscr{U}$ is spanned by exponential polynomials.*

**Proof:** By Lemma 5.2.29 there exists a zero dimensional ideal $\mathscr{I} := \mathscr{I}(\mathscr{U}) = \langle F \rangle$ such that $\mathscr{U} = \ker F(\tau)$. The common zeros $Z(\mathscr{I})$ and their associated multiplicities $\mathscr{Q}_\theta$ define $\theta$-multiplicities $\mathscr{Q}_\theta^*$, $\theta \in Z(\mathscr{I})$, hence

$$\mathscr{U} \subseteq \sum_{\theta \in Z(\mathscr{I})} \mathscr{Q}_\theta^* e_\theta,$$

and the same dimension argument as above yields equality of the two spaces. □

## 5.3 Filterbanks

We will define filterbanks in a very general setting now, using not only scalar decimations but arbitrary matrices. This will require some definitions and concepts.

### 5.3.1 Dilation matrices and the Smith factorization

**Definition 5.3.1.** A matrix $\Xi \in \mathbb{Z}^{s\times s}$ is called a SCALING MATRIX if it is nonsingular and EXPANSIVE, that is

$$\lim_{n\to\infty} \left\| \Xi^{-n} \right\| = 0 \tag{5.3.1}$$

for some matrix norm $\|\cdot\|$.

**Remark 5.3.2.** Strictly speaking, (5.3.1) means that the inverse $\Xi^{-1}$ is CONTRACTIVE; both means that all eigenvalues of $\Xi$ are $> 1$ in modulus with emphasis on the strict inequality.

**Definition 5.3.3.** A matrix $A \in \mathbb{Z}^{s\times s}$ is called UNIMODULAR if $|\det A| = 1$.

**Remark 5.3.4.** In general, a matrix $A \in R^{s\times s}$ over a ring $R$ is called unimodular if $\det R \in R^\times$ is a unit. By Cramer's rule it can be easily shown that a matrix has an inverse in $R^{s\times s}$ if and only if it is unimodular.

The following result is well-known in the theory of matrices over rings, but also under different names like *"fundamental theorem for finite groups"*, cf. [Latour et al., 1998]. A proof can be found, for example, in [Marcus and Minc, 1969].

**Theorem 5.3.5** (SMITH FACTORIZATION)**.** *For any matrix $A \in \mathbb{Z}^{s\times s}$ there exist unimodular matrices $U_1, U_2 \in \mathbb{Z}^{s\times s}$ and a diagonal matrix $\Sigma \in \mathbb{Z}^{s\times s}$ such that*

$$A = U_1 \Sigma U_2. \tag{5.3.2}$$

**Remark 5.3.6.**

1. In general neither the factors $U_1, U_2$ nor the diagonal matrix $\Sigma$ are unique.
2. There exists a SMITH NORMAL FORM that orders the diagonal elements of $\Sigma$ by divisibility and relates them to the minors of $A$, cf. [Marcus and Minc, 1969].
3. The Smith factorization can be computed efficiently and symbolically by a combination of Gaussian elimination and euclidean division with remainder of integers.
4. Since
$$\det A = \det U_1 \det \Sigma \det U_2 = \det \Sigma = \prod_{j=1}^{s} \sigma_{jj} = \prod_{j=1}^{s} \lambda_j$$
where $\lambda_j$ are the eigenvalues of $A$, there is no further relationship between the SMITH VALUES $\sigma_j := \sigma_{jj}$ of the factorization (5.3.2) and the eigenvalues of $A$, regardless of whether one considers the normal form or not.

**Lemma 5.3.7.** *If $\Xi \in \mathbb{Z}^{s\times s}$ is an expansive matrix, then*

$$\mathbb{Z}^s = \bigcup_{\xi\in E_\Xi} \left(\xi + \Xi\mathbb{Z}^s\right), \qquad E_\Xi := \Xi[0,1)^s \cap \mathbb{Z}^s \simeq \mathbb{Z}^s/\Xi\mathbb{Z}^s, \tag{5.3.3}$$

*is a decomposition of $\mathbb{Z}^s$ into $\#E_\Xi = |\det \Xi|$ equivalence classes modulo $\Xi$.*

**Proof:** We use the Smith factorization (5.3.2) and remark that for $\alpha, \beta \in \mathbb{Z}^s$ we have

$$\alpha - \beta \in \Xi\mathbb{Z}^s = U_1 \Sigma \underbrace{U_2\mathbb{Z}^s}_{=\mathbb{Z}^s} \qquad \Leftrightarrow \qquad U_1^{-1}\alpha - U_2^{-1}\beta \in \Sigma\mathbb{Z}^s,$$

hence

$$\mathbb{Z}^s/\Xi\mathbb{Z}^s \simeq \mathbb{Z}^s/\Sigma\mathbb{Z}^s = \bigotimes_{j=1}^{s} \mathbb{Z}/\sigma_j\mathbb{Z},$$

and the group on the right hand side as $\prod_j |\sigma_j| = |\det \Xi|$ elements. For $\alpha \in \mathbb{Z}^s$ we choose as a representer the element

$$\alpha + \Xi\mathbb{Z}^s \ni \xi := \alpha - \Xi\left\lfloor \Xi^{-1}\alpha \right\rfloor = \Xi\underbrace{\left(\Xi^{-1}\alpha - \left\lfloor \Xi^{-1}\alpha \right\rfloor\right)}_{\in[0,1)^s} \in \Xi[0,1)^s,$$

and the equivalence classes are disjoint since for $\xi, \xi' \in \Xi[0,1)^s$ the relationship $\xi - \xi' = \Xi\beta$ leads to

$$\beta = \Xi^{-1}\left(\xi - \xi'\right) \in \Xi^{-1}\Xi(-1,1)^s = (-1,1)^s$$

and thus to $\beta = 0$. □

**Definition 5.3.8.** The DUAL QUOTIENT GROUP of the QUOTIENT GROUP $\mathbb{Z}^s/\Xi\mathbb{Z}^s$ is $\mathbb{Z}/\Xi^T\mathbb{Z}^s$ with the representers $E'_\Xi := \Xi^T[0,1)^s$.

### 5.3.2 Fourier matrices and sampling

The use of the word "dual" in Definition 5.3.8 is purposefully chosen as the next result shows.

**Proposition 5.3.9** (Fourier matrices)**.** *We have*

$$\frac{1}{|\det \Xi|} \sum_{\xi \in E_\Xi} e^{2\pi i \xi^T \Xi^{-T} \xi'} = \delta_{\xi',0} \qquad \xi' \in E'_\Xi, \tag{5.3.4}$$

*and*

$$\frac{1}{|\det \Xi|} \sum_{\xi' \in E'_\Xi} e^{2\pi i \xi^T \Xi^{-T} \xi'} = \delta_{\xi,0} \qquad \xi \in E_\Xi, \tag{5.3.5}$$

*that is, the* FOURIER MATRIX

$$F_\Xi = \left( e^{2\pi i \xi^T \Xi^{-T} \xi'} \; : \; \begin{array}{c} \xi \in E_\Xi \\ \xi' \in E'_\Xi \end{array} \right) \in \mathbb{C}^{E_\Xi \times E'_\Xi} \tag{5.3.6}$$

*is unitary up to the factor* $|\det \Xi|$.

**Proof:** For $\xi' = 0$, (5.3.4) is obvious, otherwise we note that

$$0 \neq \xi' \in \Xi^T [0,1)^s \cap \mathbb{Z}^s = U_2^T \Sigma \underbrace{U_1^T \mathbb{Z}^s}_{=\mathbb{Z}^s} \cap U_2^T \mathbb{Z}^s = U_2^T \left( \Sigma [0,1)^s \cap \mathbb{Z}^s \right) =: U_2^T \eta', \qquad \eta' \in E_\Sigma \subset \mathbb{Z}^s,$$

and, by the same argument, also

$$\xi \in U_1 \left( \Sigma [0,1)^s \cap \mathbb{Z}^s \right) := U_1 \eta, \qquad \eta \in E_\Sigma \subset \mathbb{Z}^s.$$

Hence,

$$\begin{aligned} \sum_{\xi \in E_\Xi} e^{2\pi i \xi^T \Xi^{-T} \xi'} &= \sum_{\eta \in \Sigma [0,1)^s \cap \mathbb{Z}^s} e^{2\pi i \eta^T U_1^T \Xi^{-T} U_2^T \eta'} \\ &= \sum_{\eta \in E_\Sigma} e^{2\pi i \eta^T U_1^T U_1^{-T} \Sigma^{-1} U_2^{-T} U_2^T \eta'} = \sum_{\eta \in E_\Sigma} e^{2\pi i \eta^T \Sigma^{-1} \eta'}, \qquad \eta' \in E_\Sigma, \end{aligned}$$

so that the sum (5.3.4) depends only on $\Sigma$. Since

$$E_\Sigma = E'_\Sigma = \bigotimes_{j=1}^s \mathbb{Z}_{\sigma_j},$$

we also conclude for $\xi' \neq 0$, hence $\eta' \neq 0$, that

$$\begin{aligned} \sum_{\xi \in E_\Xi} e^{2\pi i \xi^T \Xi^{-T} \xi'} &= \sum_{j_1=0}^{\sigma_1 - 1} \cdots \sum_{j_s=0}^{\sigma_s - 1} \left( e^{2\pi i \eta'_1/\sigma_1} \right)^{j_1} \cdots \left( e^{2\pi i \eta'_s/\sigma_s} \right)^{j_s} \\ &= \sum_{j_1=0}^{\sigma_1 - 1} \left( e^{2\pi i \eta'_1/\sigma_1} \right)^{j_1} \cdots \underbrace{\sum_{j_s=0}^{\sigma_s - 1} \left( e^{2\pi i \eta'_s/\sigma_s} \right)^{j_s}}_{= \left(1 - \left( e^{2\pi i \eta'_s/\sigma_s} \right)^{\sigma_s} \right) / \left( 1 - e^{2\pi i \eta'_s/\sigma_s} \right)} = \prod_{j=1}^s \frac{1 - e^{2\pi i \eta'_s}}{1 - e^{2\pi i \eta'_s/\sigma_s}} \\ &= 0, \end{aligned}$$

which yields (5.3.4) and the same argument also leads to (5.3.5). The unimodularity of the Fourier matrix is due to

$$\begin{aligned} F_\Xi^H F_\Xi &= \left(e^{-2\pi i \xi^T \Xi^{-T} \xi'} : \begin{matrix} \xi' \in E'_\Xi \\ \xi \in E_\Xi \end{matrix}\right)\left(e^{2\pi i \xi^T \Xi^{-T} \eta'} : \begin{matrix} \xi \in E_\Xi \\ \eta' \in E'_\Xi \end{matrix}\right) \\ &= \left(\sum_{\xi \in E_\Xi} e^{-2\pi i \xi^T \Xi^{-T} \xi'} e^{2\pi i \xi^T \Xi^{-T} \eta'} : \begin{matrix} \xi' \in E_\Xi \\ \eta' \in E'_\Xi \end{matrix}\right) \\ &= \left(\sum_{\xi \in E_\Xi} e^{-2\pi i \xi^T \Xi^{-T} (\xi' - \eta')} : \begin{matrix} \xi' \in E_\Xi \\ \eta' \in E'_\Xi \end{matrix}\right) = |\det \Xi| \left(\delta_{\xi',\eta'} : \begin{matrix} \xi' \in E_\Xi \\ \eta' \in E'_\Xi \end{matrix}\right) = |\det \Xi| \, I, \end{aligned}$$

which completes the proof. □

**Definition 5.3.10** (Up- and Downsampling)**.** The DOWNSAMPLING operator is defined as

$$\downarrow_\Xi c := c(\Xi \cdot) \tag{5.3.7}$$

the UPSAMPLING operator as

$$\uparrow_\Xi c(\alpha) = \begin{cases} c\left(\Xi^{-1}\alpha\right), & \alpha \in \Xi \mathbb{Z}^s, \\ 0, & \text{otherwise}, \end{cases} \qquad \alpha \in \mathbb{Z}^s. \tag{5.3.8}$$

**Remark 5.3.11.** The way how downsampling works, namely by extracting the signal components corresponding to the equivalence class $\xi = 0$, is clear, while upsampling takes a signal and "stretches" into this class $\Xi\mathbb{Z}^s$; the other parts of the signal is filled with zero values. These operators are partial inverses:

$$\downarrow_\Xi \uparrow_\Xi = I \neq \uparrow_\Xi \downarrow_\Xi, \tag{5.3.9}$$

but we still have that

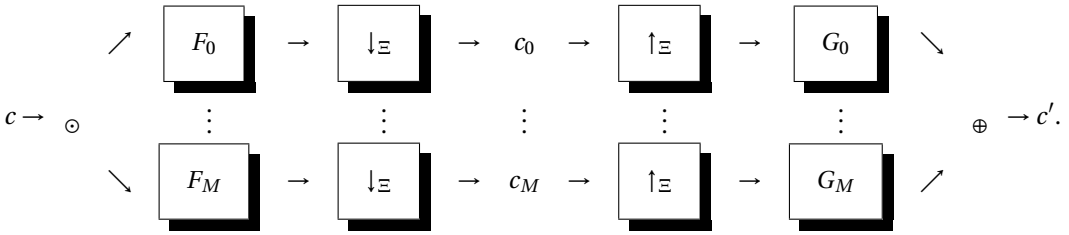
$$\sum_{\xi \in E_\Xi} \tau^\xi \uparrow_\Xi \downarrow_\Xi \tau^{-\xi} = I, \tag{5.3.10}$$

which is all the mathematics behind the so-called LAZY FILTERBANK.

**Exercise 5.3.1** Verify (5.3.10). ♢

### 5.3.3 Filterbanks in symbol calculus

Setting $M := |\det \Xi| - 1$, we can depict our filterbank as




The simplest way to build a perfect reconstruction filterbank is to index the filters as $F_\xi$, $G_\xi$, $\xi \in E_\Xi$, and to use the lazy filterbank

$$G_\xi = F_\xi^{-1} = \tau^\xi, \qquad \xi \in E_\Xi, \tag{5.3.11}$$

perfect reconstruction then follows readily from (5.3.10). Of course, this concept is not particularly exciting as it only decomposes a signal according to its parities[8] modulo $\Xi$ and then to recombine them. To get an algebraization of filterbanks, we need yet another concept.

[8] In the simplest case $s = 1$ and $\Xi = 2$ this would result in decomposing into odd and even components.

**Definition 5.3.12** (Matrix monomials)**.** Given $\Gamma \in \mathbb{Z}^{s\times s}$, written in terms of its column vectors,

$$\Gamma := \big(\gamma_1,\dots,\gamma_s\big), \qquad \gamma_j \in \mathbb{Z}^s, \quad j=1,\dots,s,$$

we define the MATRIX MONOMIAL $z^\Gamma := \big(z^{\gamma_j} : j=1,\dots,s\big)$. Moreover, the HADAMARD PRODUCT of $z, z' \in \mathbb{C}^s$ is defined as

$$z\,z' := \big(z_1 z_1', \cdots, z_s z_s'\big) \in \mathbb{C}^s. \tag{5.3.12}$$

This terminology allows us to describe the operation of upsampling and downsampling in symbol calculus.

**Lemma 5.3.13** (Up- & downsampling)**.** *For $c \in \ell_0(\mathbb{Z}^s)$ we have*

$$(\uparrow_\Xi c)^\sharp(z) \;=\; c^\sharp\big(z^\Xi\big), \tag{5.3.13}$$

$$(\downarrow_\Xi c)^\sharp\big(z^\Xi\big) \;=\; \frac{1}{|\det\Xi|}\sum_{\xi'\in E'_\Xi} c^\sharp\Big(e^{2\pi\Xi^{-T}\xi'} z\Big). \tag{5.3.14}$$

**Proof:** (5.3.13) follows immediately from

$$\begin{aligned}(\uparrow_\Xi c)^\sharp(z) &= \sum_{\alpha\in\mathbb{Z}^s} \uparrow_\Xi c(\alpha)\, z^\alpha = \sum_{\xi\in E_\Xi}\sum_{\alpha\in\mathbb{Z}^s} \underbrace{\uparrow_\Xi c(\xi+\Xi\alpha)}_{=\delta(\xi)c(\alpha)} z^{\xi+\Xi\alpha}\\ &= \sum_{\alpha\in\mathbb{Z}^s} c(\alpha)\,\underbrace{z^{\Xi\alpha}}_{=(z^\Xi)^\alpha} = c^\sharp\big(z^\Xi\big),\end{aligned}$$

while downsampling requires the Fourier duality (5.3.5):

$$\begin{aligned}(\downarrow_\Xi c)^\sharp\big(z^\Xi\big) &= \sum_{\alpha\in\mathbb{Z}^s} c(\Xi\alpha)\, z^{\Xi\alpha} = \sum_{\xi\in E_\Xi}\delta(\xi)\sum_{\alpha\in\mathbb{Z}^s} c(\xi+\Xi\alpha)\, z^{\xi+\Xi\alpha}\\
&= \sum_{\xi\in E_\Xi} \underbrace{\frac{1}{|\det\Xi|}\sum_{\xi'\in E'_\Xi} e^{2\pi i\xi^T\Xi^{-T}\xi'}}_{=\delta(\xi)} \sum_{\alpha\in\mathbb{Z}^s} c(\xi+\Xi\alpha)\, z^{\xi+\Xi\alpha}\\
&= \frac{1}{|\det\Xi|}\sum_{\xi\in E_\Xi}\sum_{\xi'\in E'_\Xi}\sum_{\alpha\in\mathbb{Z}^s} \underbrace{e^{2\pi i(\Xi\alpha)^T\Xi^{-T}\xi'}}_{=e^{2\pi i\alpha^T\xi'}=1} e^{2\pi i\xi^T\Xi^{-T}\xi'} c(\xi+\Xi\alpha)\, z^{\xi+\Xi\alpha}\\
&= \frac{1}{|\det\Xi|}\sum_{\xi\in E_\Xi}\sum_{\xi'\in E'_\Xi}\sum_{\alpha\in\mathbb{Z}^s} e^{2\pi i(\xi+\Xi\alpha)^T\Xi^{-T}\xi'} c(\xi+\Xi\alpha)\, z^{\xi+\Xi\alpha}\\
&= \frac{1}{|\det\Xi|}\sum_{\xi\in E'_\Xi}\sum_{\alpha\in\mathbb{Z}^s} e^{2\pi i\alpha^T\Xi^{-T}\xi'} c(\alpha) z^\alpha = \frac{1}{|\det\Xi|}\sum_{\xi'\in E'_\Xi}\sum_{\alpha\in\mathbb{Z}^s} c(\alpha)\Big(e^{2\pi i\Xi^{-T}\xi'} z\Big)^\alpha\\
&= \frac{1}{|\det\Xi|}\sum_{\xi'\in E'_\Xi} c^\sharp\Big(e^{2\pi\Xi^{-T}\xi'} z\Big),\end{aligned}$$

which is (5.3.14). □

The vectors $z_{\xi'} := e^{2\pi\Xi^{-T}\xi'}$, $\xi' \in E'_\Xi$, from (5.3.14) deserve a closer inspection.

**Example 5.3.14.** We consider the simplest case $s=1$.

1. For $\Xi = 2$ we obtain $E'_\Xi = \mathbb{Z}/2\mathbb{Z} = \mathbb{Z}_2 = \{0,1\}$ and numbers are $e^{\pi\xi'} = \pm 1$.
2. Getting slightly more general with $\Xi = n$, then $E'_\Xi = \mathbb{Z}_n$ and the respective values $\{e^{2\pi ik/n} : k \in \mathbb{Z}_n\}$ are just the $n$th ROOTS OF UNITY.

Roots of unity are, in principle, nothing else than signs and exactly this is the role even of $z_\xi$. If $\xi_j$ is a column of $\Xi$, $j = 1,\dots,s$, we get

$$\left(e^{2\pi i \Xi^{-T}\xi'}\right)^{\xi_j} = e^{2\pi i \xi_j^T \Xi^{-T}\xi'}$$

and therefore

$$z_{\xi'}^{\Xi} = e^{2\pi i \Xi^T \Xi^{-T}\xi'} = e^{2\pi i \xi'} = [1\dots 1] =: \mathbf{1},$$

since $\xi' \in \mathbb{Z}^s$. This can be rephrased as follows.

**Proposition 5.3.15.** *The vectors $z_{\xi'} \in \mathbb{C}^s$ are $\Xi$th roots of unity.*

With these tools we can formalize the filterbank. We first index the filters as $F_\xi$, $\xi \in E_\Xi$ which fits the philosophy of the most interesting type of filterbanks.

**Definition 5.3.16.** A filterbank is called CRITICALLY SAMPLED if $M = |\det \Xi| - 1$. In that case we index the filters conveniently as

$$F_\xi, G_\xi, \qquad \xi \in \Xi. \tag{5.3.15}$$

The processing by the analysis filters is then of the form $\downarrow_\Xi F_\xi$ and has the symbol

$$\begin{aligned}
\left(\downarrow_\Xi F_\xi c\right)^\sharp\left(z^\Xi\right) &= \frac{1}{|\det \Xi|} \sum_{\xi' \in E'_\Xi} \left(F_\xi c\right)^\sharp\left(e^{2\pi i \Xi^{-T}\xi'} z\right) \\
&= \frac{1}{|\det \Xi|} \sum_{\xi' \in E'_\Xi} f_\xi^\sharp\left(e^{-2\pi i \Xi^{-T}\xi'} z\right) c^\sharp\left(e^{-2\pi i \Xi^{-T}\xi'} z\right) \\
&= \frac{1}{|\det \Xi|} \left[f_\xi^\sharp\left(e^{2\pi i \Xi^{-T}\xi'} z\right) : \xi' \in E'_\Xi\right]^T \left[c^\sharp\left(e^{2\pi i \Xi^{-T}\xi'} z\right) : \xi' \in E'_\Xi\right].
\end{aligned}$$

Vectorizing the resulting $c_\xi = \downarrow_\Xi F_\xi$, the filterbank can be written as a matrix-vector product

$$\begin{aligned}
&\left(c_\xi^\sharp\left(z^\Xi\right) : \xi \in E_\Xi\right) \\
&= \frac{1}{|\det \Xi|} \left(f_\xi^\sharp\left(e^{2\pi i \Xi^{-T}\xi'} z\right) : \begin{matrix} \xi \in E_\Xi \\ \xi' \in E'_\Xi \end{matrix}\right) \left(c^\sharp\left(e^{2\pi i \Xi^{-T}\xi'} z\right) : \xi' \in E'_\Xi\right)
\end{aligned} \tag{5.3.16}$$

Let us recall the names of the objects appearing in (5.3.16).

**Definition 5.3.17** (Polyphase & modulation)**.**

1. The POLYPHASE REPRESENTATION or the POLYPHASE VECTOR of a signal $c \in \ell(\mathbb{Z}^s)$ is defined as
$$\boldsymbol{c}_p^\sharp(z) := \left(c^\sharp\left(e^{2\pi i \Xi^{-T}\xi'} z\right) : \xi' \in E'_\Xi\right) \tag{5.3.17}$$

2. The MODULATION MATRIX of an analysis filterbank $F = \left(F_\xi : \xi \in E_\Xi\right)$ is
$$F(z) := \frac{1}{|\det \Xi|} \left(f_\xi^\sharp\left(e^{2\pi i \Xi^{-T}\xi'} z\right) : \begin{matrix} \xi \in E_\Xi \\ \xi' \in E'_\Xi \end{matrix}\right) \in \Lambda^{E_\Xi \times E'_\Xi}. \tag{5.3.18}$$

3. The matrix-vector representation of the analysis filterbank is then
$$\boldsymbol{c}^\sharp\left(z^\Xi\right) := \left(c_\xi^\sharp\left(z^\Xi\right) : \xi \in E_\Xi\right) = F(z)\, \boldsymbol{c}_p^\sharp(z). \tag{5.3.19}$$

**Remark 5.3.18.** In the univariate case, see (1.2.21), we built the polyphase vectors with respect to the $z$-transform, here we use symbol calculus. This is really a matter of taste and does not really make a difference.

The synthesis filterbank is more easily computed since a direct application of (5.3.13) yields

$$d^\sharp(z) = \sum_{\xi\in E_\Xi} \left(G_\xi \uparrow_\Xi c_\xi\right)^\sharp(z) = \sum_{\xi\in E_\Xi} g_\xi^\sharp(z)\, c_\xi^\sharp\left(z^\Xi\right) \tag{5.3.20}$$

and fortunately the analysis filterbank precisely yields the vector $\boldsymbol{c}^\sharp\left(z^\Xi\right)$ needed there. Since (5.3.19), on the other hand, is based on a polyphase vector, we can write the return value $c'$ in polyphase form as well and thus get for $\xi' \in E'_\Xi$ that

$$\begin{aligned}\left(\boldsymbol{c}'_p(z)\right)_{\xi'} &= d^\sharp\left(e^{2\pi i\Xi^{-T}\xi'}\, z\right) = \sum_{\xi\in E_\Xi} g_\xi^\sharp\left(e^{2\pi i\Xi^{-T}\xi'}\, z\right) c_\xi^\sharp\Big(\underbrace{e^{2\pi i\Xi^T\Xi^{-T}\xi'}}_{=1} z^\Xi\Big) \\ &= \left(\left(g_\xi^\sharp\left(e^{2\pi i\Xi^{-T}\xi'}\, z\right) : \begin{array}{c}\xi'\in E'_\Xi\\ \xi\in E_\Xi\end{array}\right)\boldsymbol{c}\left(z^\Xi\right)\right)_{\xi'},\end{aligned}$$

that is

$$\boldsymbol{c}'_p(z) = |\det\Xi|\; G(z)^T\, \boldsymbol{c}\left(z^\Xi\right) = |\det\Xi|\; G(z)^T\, F(z)\, \boldsymbol{c}_p(z), \tag{5.3.21}$$

now with the modulation matrix for $\left(g_\xi : \xi\in E_\Xi\right)$. The identity (5.3.21) immediately allows us to describe perfect reconstruction in terms of modulation matrices.

**Theorem 5.3.19** (Perfect reconstruction)**.** *A filterbank provides perfect reconstruction if and only if*

$$G^T(z)\, F(z) = \frac{1}{|\det\Xi|}\, I. \tag{5.3.22}$$

Thus a given analysis filterbank or synthesis filterbank can be completed to a perfect reconstruction filterbank if the given filterbank has an inverse[9] in $\Lambda^{E'_\Xi\times E_\Xi}$ or $\Lambda^{E_\Xi\times E'_\Xi}$, respectively. Taking into account that matrices over rings are invertible in the ring if and only if they are unimodular, we can draw the following conclusion.

**Corollary 5.3.20.** *An analysis filterbank $F$ or a synthesis filterbank $G$ can be completed to a perfect reconstruction filterbank if and only if $F(z)$ or $G(z)$ are* UNIMODULAR*, respectively.*

In other words: if $F(z)\in\Lambda^{E'_\Xi\times E_\Xi}$ is a unimodular matrix, then the filterbank can be completed by $G(z) = F^{-T}(z)$, which requires symbolic inversion[10] and transposition[11]. But this process does not directly guarantee that $F^{-T}(z)$ really has the structure of a modulation matrix. This requires some extra effort.

**Proposition 5.3.21** (Inverses of modulation matrices)**.** *If $F\in\Lambda^{s\times s}$ is unimodular, then there exist Laurent polynomials $g_\xi\in\Lambda$, $\xi\in E_\Xi$, such that*

$$F^{-1}(z) = \left(g_\xi^\sharp\left(e^{2\pi i\Xi^{-T}\xi'}\, z\right) : \begin{array}{c}\xi'\in E'_\Xi\\ \xi\in E_\Xi\end{array}\right) \in \Lambda^{E'_\Xi\times E_\Xi}. \tag{5.3.23}$$

[9] Keep in mind that the inverse of a matrix in $R^{X\times Y}$ is a matrix in $R^{Y\times X}$.
[10] Cramer's rule is a possibility though not very efficient.
[11] Which is really easy.

**Proof:** We define $g^\sharp_\xi(z) := \big(F^{-1}(z)\big)_{0,\xi}$ and remark that [12]

$$\delta_{0,\xi'} = \big(F^{-1}(z)\,F(z)\big)_{0,\xi'} = \frac{1}{|\det \Xi|} \sum_{\xi\in E_\Xi} g^\sharp_\xi(z)\, f^\sharp_\xi\Big(e^{2\pi i \Xi^{-T}\xi'} z\Big), \qquad \xi' \in E'_\Xi. \tag{5.3.24}$$

Setting $\xi' = 0$ in (5.3.24) and replacing $z$ by $e^{2\pi i\Xi^{-T}\eta'} z$ for some $\eta' \in E'_\Xi$, we get

$$1 = \frac{1}{|\det \Xi|} \sum_{\xi\in E_\Xi} g^\sharp_\xi\Big(e^{2\pi i\Xi^{-T}\eta'} z\Big)\, f^\sharp_\xi\Big(e^{2\pi i\Xi^{-T}\eta'} z\Big) = \big(G^T(z)F(z)\big)_{\eta',\eta'}. \tag{5.3.25}$$

The same trick with $\xi' \neq 0$ yields

$$0 = \frac{1}{|\det \Xi|} \sum_{\xi\in E_\Xi} g^\sharp_\xi\Big(e^{2\pi i\Xi^{-T}\eta'} z\Big)\, f^\sharp_\xi\Big(e^{2\pi i\Xi^{-T}(\xi'+\eta')} z\Big) = \big(G^T(z)F(z)\big)_{\eta',\xi'+\eta'}, \tag{5.3.26}$$

and since $\{\eta' + \xi' : \xi' \neq 0\} = E'_\Xi \setminus \{\eta'\}$, at least modulo $\Xi$, we can combine (5.3.25) and (5.3.26) into

$$\delta_{\xi',\eta'} = \big(G^T(z)F(z)\big)_{\xi',\eta'}, \qquad \text{i.e.,} \qquad G(z) = F^{-T}(z),$$

and since the inverse of a matrix is unique once it exist, we can conclude that $F^{-T}$ has the structure of a modulation matrix. □

### 5.3.4 Matrix completion and interpolatory sequences

Now we get to the interesting question of how much we have to know about a filterbank in order to be able to complete it to a perfect reconstruction one. We take the point of view that we start with the synthesis filterbank and want to construct an appropriate analysis filterbank for it. To that end, we denote by

$$\boldsymbol{g}^\sharp_p := \Big(g^\sharp_0\Big(e^{2\pi i\Xi^{-T}\xi'}\cdot\Big) : \xi' \in E'_\Xi\Big) \tag{5.3.27}$$

the polyphase vector for the filter $G_0$, which is also the 0th row of the matrix $G(z)$.

**Proposition 5.3.22.** *If the modulation matrix $G(z)$ is unimodular then $\boldsymbol{g}^\sharp_p$ is* UNIMODULAR *in the sense that*

$$1 \in \Big\langle \boldsymbol{g}^\sharp_p \Big\rangle = \Big\langle g^\sharp_0\Big(e^{2\pi i\Xi^{-T}\xi'}\cdot\Big) : \xi' \in E'_\Xi \Big\rangle. \tag{5.3.28}$$

**Proof:** Since $G$ is unimodular, there exists for any vector $\boldsymbol{y} \in \Lambda^{E_\Xi}$ a vector $\boldsymbol{a} = \big(a_{\xi'} : \xi \in E'_\Xi\big) \in \Lambda^{E'_\Xi}$ such that

$$\boldsymbol{y}(z) = G(z)\boldsymbol{a}(z), \qquad \text{namely,} \qquad \boldsymbol{a}(z) = G^{-1}(z)\boldsymbol{y}(z).$$

Choosing $y_0 = 1$ leads to

$$1 = \sum_{\xi'\in E'_\Xi} g^\sharp_0\Big(e^{2\pi i\Xi^{-T}\xi'}\cdot\Big)\, a_{\xi'},$$

which is the representation requested by (5.3.28). □

**Remark 5.3.23.** Note that (5.3.28), i.e., unimodularity of the polyphase vector, can also be expressed as the fact that the components of the polyphase vector have *no common zeros*. This directly connects to Theorem 1.2.17.

[12] Concerning the notation: $F^{-1}F \in \mathbb{C}^{E'_\Xi \times E'_\Xi}$ is an identity matrix indexed with $\xi'$.

The converse of Proposition 5.3.22 could be concluded from the QUILLEN-SUSLIN THEOREM, cf. [Cox et al., 1998] and [Park, 1995, Park and Woodburn, 1995] for an explicit algorithm.

We will give a different, more elementary proof here which, however, does not yield an immediate completion algorithm, only after prefiltering. To that end, we first observe that the representation of (5.3.28) can also be done in a special form.

**Lemma 5.3.24.** *If* (5.3.28) *holds true then there exists* $h \in \Lambda$ *such that*

$$1 = \sum_{\xi' \in E'_\Xi} h^\sharp\left(e^{2\pi i \Xi^{-T}\xi'}\cdot\right) g_0^\sharp\left(e^{2\pi i \Xi^{-T}\xi'}\cdot\right) = \sum_{\xi' \in E'_\Xi} \left(h * g_0\right)^\sharp\left(e^{2\pi i \Xi^{-T}\xi'}\cdot\right). \tag{5.3.29}$$

**Proof:** In the identity

$$1 = \sum_{\xi' \in E'_\Xi} g_0^\sharp\left(e^{2\pi i \Xi^{-T}\xi'} z\right) a_{\xi'}(z)$$

we replace $z$ by $e^{2\pi i \Xi^{-T}\eta} z$ for some $\eta \in E'_\Xi$ and note that

$$1 = \sum_{\xi' \in E'_\Xi} g_0^\sharp\left(e^{2\pi i \Xi^{-T}(\xi'+\eta)} z\right) a_{\xi'}(e^{2\pi i \Xi^{-T}\eta} z) = \sum_{\xi' \in E'_\Xi} g_0^\sharp\left(e^{2\pi i \Xi^{-T}\xi'} z\right) a_{\xi'-\eta}(e^{2\pi i \Xi^{-T}\eta} z) \tag{5.3.30}$$

due to the group structure of $E'_\Xi$. Averaging (5.3.30) over $\eta$ then yields that

$$\begin{aligned}
1 &= \frac{1}{|\det \Xi|} \sum_{\eta \in E'_\Xi} \sum_{\xi' \in E'_\Xi} g_0^\sharp\left(e^{2\pi i \Xi^{-T}\xi'} z\right) a_{\xi'-\eta}(e^{2\pi i \Xi^{-T}\eta} z) \\
&= \frac{1}{|\det \Xi|} \sum_{\xi' \in E'_\Xi} g_0^\sharp\left(e^{2\pi i \Xi^{-T}\xi'} z\right) \sum_{\eta \in E'_\Xi} a_{\xi'-\eta}(e^{2\pi i \Xi^{-T}\eta} z) \\
&= \frac{1}{|\det \Xi|} \sum_{\xi' \in E'_\Xi} g_0^\sharp\left(e^{2\pi i \Xi^{-T}\xi'} z\right) \sum_{\eta \in E'_\Xi} a_{\eta}(e^{2\pi i \Xi^{-T}(\xi'-\eta)} z) \\
&= \sum_{\xi' \in E'_\Xi} g_0^\sharp\left(e^{2\pi i \Xi^{-T}\xi'} z\right) \left(\frac{1}{|\det \Xi|} \sum_{\eta \in E'_\Xi} a_\eta(e^{-2\pi i \Xi^{-T}\eta}\cdot)\right)\left(e^{2\pi i \Xi^{-T}\xi'} z\right)
\end{aligned}$$

and therefore

$$h^\sharp := \frac{1}{|\det \Xi|} \sum_{\eta \in E'_\Xi} a_\eta(e^{-2\pi i \Xi^{-T}\eta}\cdot).$$

satisfies (5.3.29). □

**Remark 5.3.25.** Note that $h$ from (5.3.29) also has a unimodular polyphase vector.

The next concept is traditional and useful in subdivision theory when the meaning of the notion will become clear.

**Definition 5.3.26.** The SUBSYMBOL of a sequence $a \in \ell_0(\mathbb{Z}^s)$ is the LAURENT POLYNOMIAL

$$a_\xi^\sharp(z) := \sum_{\alpha \in \mathbb{Z}^s} a\left(\Xi\alpha + \xi\right) z^\alpha, \tag{5.3.31}$$

and $a$ is called INTERPOLATORY with respect to $\Xi$ if $a(\Xi\cdot) = \delta$.

The SUBSYMBOL REPRESENTATION of a symbol follows readily from the decomposition (5.3.3) and takes the form

$$a^\sharp(z) = \sum_{\alpha \in \mathbb{Z}^s} a(\alpha)\, z^\alpha = \sum_{\xi \in E_\Xi} \sum_{\alpha \in \mathbb{Z}^s} a(\Xi\alpha + \xi)\, z^{\Xi\alpha+\xi} = \sum_{\xi \in E_\Xi} z^\xi a_\xi^\sharp\left(z^\Xi\right). \tag{5.3.32}$$

The next result is well-known from subdivision theory and describes interpolatory filters.

**Lemma 5.3.27.** *A sequence $a \in \ell_0(\mathbb{Z}^s)$ is interpolatory if and only if*

$$\frac{1}{|\det \Xi|} \sum_{\xi' \in E'_\Xi} a^\sharp\left(e^{2\pi i \Xi^{-T}\xi'} \cdot\right) = 1. \tag{5.3.33}$$

**Proof:** Substituting the subsymbol decomposition (5.3.32) into the left hand side of (5.3.33), we get that

$$\begin{aligned}
&\frac{1}{|\det \Xi|} \sum_{\xi' \in E'_\Xi} a^\sharp\left(e^{2\pi i \Xi^{-T}\xi'} z\right) \\
&= \frac{1}{|\det \Xi|} \sum_{\xi' \in E'_\Xi} \sum_{\xi \in E_\Xi} \left(e^{2\pi i \Xi^{-T}\xi'} z\right)^\xi a_\xi^\sharp\left(e^{2\pi i \Xi^T \Xi^{-T}\xi'} z^\Xi\right) \\
&= \frac{1}{|\det \Xi|} \sum_{\xi \in E_\Xi} z^\xi a_\xi^\sharp\left(z^\Xi\right) \underbrace{\sum_{\xi' \in E'_\Xi} e^{2\pi i \xi^T \Xi^{-T}\xi'}}_{=\delta_{\xi,0}|\det \Xi|} = a_0^\sharp\left(z^\Xi\right) = \sum_{\alpha \in \mathbb{Z}^s} \underbrace{a(\Xi\alpha)}_{=\delta_{\alpha,0}} z^{\Xi\alpha} = 1,
\end{aligned}$$

as claimed. □

**Theorem 5.3.28.** *The polyphase vector $\boldsymbol{g}_p^\sharp$ from* (5.3.27) *is unimodular if and only if there exist $h \in \ell_0(\mathbb{Z}^s)$ such that $h * g_0$ is interpolatory.*

**Proof:** The direction "⇒" is the above construction, while the fact that $h * g_0$ is interpolatory yields (5.3.29) and therefore that $1 \in \left\langle \boldsymbol{g}_p^\sharp \right\rangle$, hence "⇐". □

The advantage of interpolatory sequences is that they define filters that admit a simple unimodular completion, due to the fact that the coefficients for the representation of 1 in (5.3.31) are particularly simple, namely $|\det \Xi|^{-1}$.

**Theorem 5.3.29.** *If $a \in \ell_0(\mathbb{Z}^s)$ is interpolatory, then the matrix*

$$G(z) = \begin{pmatrix} \left(e^{-2\pi \Xi^{-T}\xi'} z\right)^\xi & : & \xi \in E_\Xi \setminus \{0\} \\ \hline a^\sharp\left(e^{-2\pi i \Xi^{-T}\xi'} z\right) & : & \xi' \in E'_\Xi \end{pmatrix} \tag{5.3.34}$$

*is unimodular in $\Lambda$.*

**Proof:** Writing $a^\sharp$ in its subsymbol decomposition modulo $\Xi$

$$a^\sharp(z) = a_0^\sharp(z) + \sum_{\xi \in E_\Xi \setminus \{0\}} z^\xi a_\xi^\sharp\left(z^\Xi\right) = 1 + \sum_{\xi \in E_\Xi \setminus \{0\}} z^\xi a_\xi^\sharp\left(z^\Xi\right),$$

we obtain for $\xi' \in E'_\Xi$ that

$$\begin{aligned}
a^\sharp\left(e^{2\pi i \Xi^{-T}\xi'} z\right) &= 1 + \sum_{\xi \in E_\Xi \setminus \{0\}} \left(e^{2\pi i \Xi^{-T}\xi'} z\right)^\xi a_\xi^\sharp\left(e^{2\pi i \Xi^T \Xi^{-T}\xi'} z^\Xi\right) \\
&= 1 + \sum_{\xi \in E_\Xi \setminus \{0\}} \left(e^{2\pi i \Xi^{-T}\xi'} z\right)^\xi a_\xi^\sharp\left(z^\Xi\right),
\end{aligned}$$

hence

$$1 = a^\sharp\left(e^{2\pi i \Xi^{-T}\xi'} z\right) - \sum_{\xi \in E_\Xi \setminus \{0\}} \left(e^{2\pi i \Xi^{-T}\xi'} z\right)^\xi a_\xi^\sharp\left(z^\Xi\right), \qquad \xi' \in E'_\Xi$$

and thus

$$\left(1,\dots,1\right) = \left(\left(-a^\sharp_\xi\left(z^\Xi\right) : \xi \in E_\Xi \setminus \{0\}\right) \quad 1\right) G(z). \tag{5.3.35}$$

This leads to

$$A(z) := \left(\begin{array}{ccc} \left(e^{2\pi\Xi^{-T}\xi'}z\right)^\xi & : & \xi \in E_\Xi \setminus \{0\} \\ \hline 1 & : & \xi' \in E'_\Xi \end{array}\right) = \begin{pmatrix} I & & 0 \\ -a^\sharp_\xi\left(z^\Xi\right) : \xi \in E_\Xi \setminus \{0\} & & 1 \end{pmatrix} G(z). \tag{5.3.36}$$

The matrix $A(z)$ can further be decomposed as

$$A(z) = \operatorname{diag}\left(z^\xi : \xi \in E_\Xi\right)\left(e^{2\pi i\xi^T\Xi^{-T}\xi'} \quad : \quad \begin{array}{c}\xi \in E_\Xi \\ \xi' \in E'_\Xi\end{array}\right) =: D_\Xi(z)\,F_\Xi,$$

where $F_\Xi$ is the Fourier matrix from (5.3.6). Therefore,

$$G(z) = \begin{pmatrix} I & & 0 \\ -a^\sharp_\xi\left(z^\Xi\right) : \xi \in E_\Xi \setminus \{0\} & & 1 \end{pmatrix}^{-1} D_\Xi(z)F_\Xi = \begin{pmatrix} I & & 0 \\ a^\sharp_\xi\left(z^\Xi\right) : \xi \in E_\Xi \setminus \{0\} & & 1 \end{pmatrix} D_\Xi(z)F_\Xi \tag{5.3.37}$$

and

$$\det G(z) = \det F_\Xi \det D_\Xi(z) = \pm\sqrt{|\det\Xi|}\prod_{\xi\in E_\Xi} z^\xi \in \Lambda^*$$

which proves unimodularity. □

**Exercise 5.3.2** Show that

$$\begin{pmatrix} I & 0 \\ v^T & 1 \end{pmatrix}^{-1} = \begin{pmatrix} I & 0 \\ -v^T & 1 \end{pmatrix}$$

♢

**Corollary 5.3.30.** *The matrices*

$$D_\Xi^{-1}(z)G(z) = \frac{1}{\det F_\Xi}\left(\begin{array}{ccc} e^{-2\pi\xi^T\Xi^{-T}\xi'} & : & \xi \in E_\Xi \setminus \{0\} \\ & : & \xi' \in E'_\Xi \\ \hline a^\sharp\left(e^{-2\pi i\Xi^{-T}\xi'}z\right) & : & \xi' \in E'_\Xi \end{array}\right)$$

*and*

$$G(z)D_\Xi^{-1}(z) = \frac{1}{\det F_\Xi}\left(\begin{array}{ccc} e^{-2\pi\xi^T\Xi^{-T}\xi'}z^{\xi-\xi'} & : & \xi \in E_\Xi \setminus \{0\} \\ & : & \xi' \in E'_\Xi \\ \hline a^\sharp\left(e^{-2\pi i\Xi^{-T}\xi'}z\right) & : & \xi' \in E'_\Xi \end{array}\right)$$

*have determinant* 1.

**Proposition 5.3.31.** *The polyphase vector and the subsymbol vector are related by*

$$\left(a^\sharp\left(e^{2\pi i\Xi^{-T}\xi'}z\right) : \xi' \in E'_\Xi\right) = F_\Xi^T D_\Xi(z)\left(a^\sharp_\xi\left(z^\Xi\right) : \xi \in E_\Xi\right) \tag{5.3.38}$$

*and one of them is unimodular if and only if the other one is unimodular.*

**Proof:** Substituting $z = e^{2\pi i\Xi^{-T}\xi'}z$, $\xi' \in E'_\Xi$ in the subsymbol decomposition (5.3.32) we get that

$$a^\sharp\left(e^{2\pi i\Xi^{-T}\xi'}z\right) = \sum_{\xi\in E_\Xi} e^{2\pi i\xi^T\Xi^{-T}\xi'}z^\xi a^\sharp_\xi\left(e^{2\pi i\Xi^T\Xi^{-T}\xi'}z^\Xi\right) = e_{\xi'}F_\Xi^T D_\Xi(z)\left(a^\sharp_\xi\left(z^\Xi\right) : \xi \in E_\Xi\right),$$

which is (5.3.38). Moreover, if $h \in \Lambda^{E'_\Xi}$ is such that

$$1 = \left(h_{\xi'} : \xi' \in E'_\Xi\right)^T \left(a^\sharp\left(e^{2\pi i \Xi^{-T}\xi'} z\right) : \xi' \in E'_\Xi\right) = \left(h_{\xi'} : \xi' \in E'_\Xi\right)^T F_\Xi^T D_\Xi(z) \left(a_\xi^\sharp\left(z^\Xi\right) : \xi \in E_\Xi\right),$$

then

$$\tilde{h} = D_\Xi(z) F_\Xi h$$

obviously is a coefficient vector for the combination of 1. The converse works the same way. □

# Bibliography

[Andoyer, 1906] Andoyer, H. (1906). Interpolation. In *Encyclopédie de Sciences Mathematiques, Tome I, Vol. 4*. Gauthier–Villars.

[Basu et al., 2003] Basu, S., Pollack, R., and Roy, M.-F. (2003). *Algorithms in Real Algebraic Geometry*, volume 10 of *Algorithms and Computation in Mathematics*. Springer.

[Bauschinger, 1900] Bauschinger, J. (1900). Interpolation. In *Encyklopädie der Mathematischen Wissenschaften, Bd. I, Teil 2*, pages 800–821. B. G. Teubner, Leipzig.

[Birkhoff, 1979] Birkhoff, G. (1979). The algebra of multivariate interpolation. In Coffman, C. and Fix, G., editors, *Constructive Approaches to Mathematical Models*, pages 345–363. Academic Press Inc.

[Boor, 1994] Boor, C. d. (1994). Gauss elimination by segments and multivariate polynomial interpolation. In Zahar, R. V. M., editor, *Approximation and Computation: A Festschrift in Honor of Walter Gautschi*, pages 87–96. Birkhäuser Verlag.

[Boor, 2005] Boor, C. d. (2005). Divided differences. *Surveys in Approximation Theory*, 1:46–69. [Online article at] http://www.math.technion.ac.il/sat.

[Boor and Ron, 1990] Boor, C. d. and Ron, A. (1990). On multivariate polynomial interpolation. *Constr. Approx.*, 6:287–302.

[Boor and Ron, 1991] Boor, C. d. and Ron, A. (1991). On polynomial ideals of finite codimension with applications to box spline theory. *J. Math. Anal. and Appl.*, 158:168–193.

[Boor and Ron, 1992] Boor, C. d. and Ron, A. (1992). The least solution for the polynomial interpolation problem. *Math. Z.*, 210:347–378.

[Borchardt, 1860] Borchardt, W. (1860). Über eine Interpolationsformel für eine Art symmetrischer Funktionen und deren Anwendung. *Abh. d. Preuß. Akad. d. Wiss.*, pages 1–20.

[Bos et al., 2007] Bos, L., de Marchi, S., Vianello, M., and Xu, Y. (2007). Bivariate Lagrange interpolation at the Padua points: the ideal theory approach. *Numer. Math.*, 108:43–57.

[Buchberger, 1965] Buchberger, B. (1965). *Ein Algorithmus zum Auffinden der Basiselemente des Restklassenrings nach einem nulldimensionalen Polynomideal*. PhD thesis, Innsbruck.

[Buchberger, 1985] Buchberger, B. (1985). Gröbner bases: An algorithmic method in polynomial ideal theory. In Bose, N. K., editor, *Multidimensional Systems Theory*, pages 184–232. D. Reidel Publishing Company.

[Buchberger, 1998] Buchberger, B. (1998). An introduction to gröbner bases. In Buchberger, B. and Winkler, F., editors, *Groebner Bases and Applications (Proc. of the Conf. 33 Years of Groebner Bases)*, volume 251 of *London Math. Soc. Lecture Notes*, pages 3–31. Cambridge University Press. to appear.

[Busch, 1990] Busch, J. R. (1990). A note on lagrange interpolation in $\mathbb{R}^2$. *Rev. Union Matem. Argent.*, 36:33–38.

[Caliari et al., 2006] Caliari, M., Bos, L., de Marchi, S., Vianello, M., and Xu, Y. (2006). Bivariate Lagrange interpolation at the Padua points: the generating curve approach. *J. Approx. Theory*, 143:15–25.

[Calvetti et al., 2003] Calvetti, D., Reichel, L., and Sgallari, F. (2003). A modified companion matrix method based on Newton polynomials. In Olshevsky, V., editor, *Fast Algorithms for Structured Matrices: Theory and Applications,*, volume 323 of *Contemporary Mathematics*. Amer. Math. Soc.

[Carnicer and Sauer, 2018] Carnicer, J. and Sauer, T. (2018). Observations on interpolation by total degree polynomials in two variables. *Constr. Approx.*, 47:373–389. arXiv:1610.01850.

[Cavaretta et al., 1991] Cavaretta, A. S., Dahmen, W., and Micchelli, C. A. (1991). *Stationary Subdivision*, volume 93 (453) of *Memoirs of the AMS*. Amer. Math. Soc.

[Chung and Yao, 1977] Chung, K. C. and Yao, T. H. (1977). On lattices admitting unique Lagrange interpolation. *SIAM J. Num. Anal.*, 14:735–743.

[Cohen et al., 1999] Cohen, A. M., Cuypers, H., and Sterk, M., editors (1999). *Some Tapas of Computer Algebra*, volume 4 of *Algorithms and Computations in Mathematics*. Springer.

[Corless et al., 2004] Corless, R. M., Watt, S. M., and Zhi, L. (2004). *QR* factoring to compute the gcd of univariate approximate polynomials. *IEEE Trans. Sig. Proc.*, 52:3394–3402.

[Cox et al., 1996] Cox, D., Little, J., and O'Shea, D. (1996). *Ideals, Varieties and Algorithms*. Undergraduate Texts in Mathematics. Springer–Verlag, 2. edition.

[Cox et al., 1998] Cox, D., Little, J., and O'Shea, D. (1998). *Using Algebraic Geometry*, volume 185 of *Graduate Texts in Mathematics*. Springer Verlag.

[DeVilliers et al., 2000] DeVilliers, J. M., Micchelli, C. A., and Sauer, T. (2000). Building refinable functions from their values at integers. *Calcolo*, 37(3):139–158.

[Eisenbud, 1994] Eisenbud, D. (1994). *Commutative Algebra with a View Toward Algebraic Geometry*, volume 150 of *Graduate Texts in Mathematics*. Springer.

[Farouki and Rajan, 1987] Farouki, R. T. and Rajan, V. T. (1987). On the numerical condition of polynomials in Bernstein form. *Comput. Aided Geom. Design*, 4:191–216.

[Fieldsteel and Schenck, 2017] Fieldsteel, N. and Schenck, H. (2017). Polynomial interpolation in higher dimension: From simplicial complexes to *GC* sets. *SIAM J. Numer. Anal.*, 55:131–143.

[Fischer, 1984] Fischer, G. (1984). *Lineare Algebra*. Vieweg.

[Föllinger, 2000] Föllinger, O. (2000). *Laplace-, Fourier- und z–Transformation*. Hüthig.

[Freitag and Busam, 2005] Freitag, E. and Busam, R. (2005). *Complex Analysis*. Springer.

[Gasca and Maeztu, 1982] Gasca, M. and Maeztu, J. I. (1982). On Lagrange and Hermite interpolation in $\mathbb{R}^k$. *Numer. Math.*, 39:1–14.

[Gasca and Sauer, 2000a] Gasca, M. and Sauer, T. (2000a). On bivariate Hermite interpolation with minimal degree polynomials. *SIAM J. Numer. Anal.*, 37:772–798.

[Gasca and Sauer, 2000b] Gasca, M. and Sauer, T. (2000b). On the history of multivariate polynomial interpolation. *J. Comput. Appl. Math.*, 122:23–35.

[Gasca and Sauer, 2000c] Gasca, M. and Sauer, T. (2000c). Polynomial interpolation in several variables. *Advances Comput. Math.*, 12:377–410.

[Gathen and Gerhard, 1999] Gathen, J. v. z. and Gerhard, J. (1999). *Modern Computer Algebra*. Cambridge University Press.

[Gauss, 1816] Gauss, C. F. (1816). Methodus nova integralium valores per approximationem inveniendi. *Commentationes societate regiae scientiarum Gottingensis recentiores*, III.

[Gautschi, 1997] Gautschi, W. (1997). *Numerical Analysis. An Introduction*. Birkhäuser.

[Goldberg, 1958] Goldberg, S. (1958). *Introduction to Difference Equations*. John Wiley & Sons. Dover reprint 1986.

[Golub and van Loan, 1996] Golub, G. and van Loan, C. F. (1996). *Matrix Computations*. The Johns Hopkins University Press, 3rd edition.

[González-Vega et al., 1999] González-Vega, L., Rouillier, F., Roy, M.-F., and Trujillo, G. (1999). Symbolic recipes for polynomial system solving. In Cohen, A. M., Cuypers, H., and Sterk, M., editors, *Some Tapas in Computer Algebra*, volume 4 of *Algorithms and Computations in Mathematics*, chapter 2, pages 34–65. Springer.

[Gould, 1971] Gould, H. W. (1971). Noch einmal die Stirlingschen Zahlen. *Jber. Deutsch. Math.-Verein*, 73:149–152.

[Graham et al., 1998] Graham, R. L., Knuth, D. E., and Patashnik, O. (1998). *Concrete Mathematics*. Addison–Wesley, 2nd edition.

[Gröbner, 1937] Gröbner, W. (1937). Über das Macaulaysche inverse System und dessen Bedeutung für die Theorie der linearen Differentialgleichungen mit konstanten Koeffizienten. *Abh. Math. Sem. Hamburg*, 12:127–132.

[Gröbner, 1939] Gröbner, W. (1939). Über die algebraischen Eigenschaften der Integrale von linearen Differentialgleichungen mit konstanten Koeffizienten. *Monatsh. Math.*, 47:247–284.

[Gröbner, 1968] Gröbner, W. (1968). *Algebraische Geometrie I*. Number 273 in B.I–Hochschultaschenbücher. Bibliographisches Institut Mannheim.

[Gröbner, 1970] Gröbner, W. (1970). *Algebraische Geometrie II*. Number 737 in B.I–Hochschultaschenbücher. Bibliographisches Institut Mannheim.

[Hakopian et al., 2009] Hakopian, H., Jetter, K., and Zimmermann, G. (2009). A new proof of the Gasca-Maeztu conjecture for $n = 4$. *J. Approx. Theory*, 159:224–242.

[Hakopian et al., 2014] Hakopian, H., Jetter, K., and Zimmermann, G. (2014). The Gasca-Maeztu conjecture for $n = 5$. *Numer. Math.*, 127:685–713.

[Hamming, 1989] Hamming, R. W. (1989). *Digital Filters*. Prentice–Hall. Republished by Dover Publications, 1998.

[Higham, 2002] Higham, N. J. (2002). *Accuracy and stability of numerical algorithms*. SIAM, 2nd edition.

[Hille, 1982] Hille, E. (1982). *Analytic Function Theory*. Chelsea Publishing Company, 2nd edition.

[Isaacson and Keller, 1966] Isaacson, E. and Keller, H. B. (1966). *Analysis of Numerical Methods*. John Wiley & Sons.

[Jacobi, 1835] Jacobi, C. G. J. (1835). Theoremata nova algebraica circa systema duarum aequationum inter duas variabiles propositarum. *Crelle J. reine und angew. Math.*, 14:281–288.

[Jordan, 1965] Jordan, C. (1965). *Calculus of finite differences*. Chelsea, 3rd edition.

[Kobbelt, 2000] Kobbelt, L. (2000). $\sqrt{3}$–subdivision. In *Proceedings of SIGGRAPH 2000*, pages 103–112.

[Kronecker, 1866] Kronecker, L. (1866). Über einige Interpolationsformeln für ganze Funktionen mehrerer Variabeln. *Monatsberichte der Königlich Preußsischen Akademie der Wissenschaften zu Berlin, 1865*, pages 686–691. Lecture at the Academy of Sciences, December 21, 1865.

[Laidacker, 1969] Laidacker, M. A. (1969). Another theorem relating Sylvester's matrix and the greatest common divisor. *Math. Mag.*, 42:126–128.

[Latour et al., 1998] Latour, V., Müller, J., and Nickel, W. (1998). Stationary subdivision for general scaling matrices. *Math. Z.*, 227:645–661.

[Lee and Phillips, 1988] Lee, S. L. and Phillips, G. M. (1988). Polynomial interpolation at points of a geometric mesh on a triangle. *Proc. Roy. Soc. Edinburgh*, 108A:75–87.

[Lorentz, 1966] Lorentz, G. G. (1966). *Approximation of functions*. Chelsea Publishing Company.

[Lorentz, 2000] Lorentz, R. A. (2000). Multivariate Hermite interpolation by algebraic polynomials: a survey. *J. Comput. Appl. Math.*, 122:167–201. Numerical analysis 2000, Vol. II: Interpolation and extrapolation.

[Lubich, 2008] Lubich, C. (2008). *From Quantum to Classical Molecular Dynamics: Reduced Models and Numerical Analysis*. European Mathematical Society.

[Macaulay, 1916] Macaulay, F. S. (1916). *The Algebraic Theory of Modular Systems*. Number 19 in Cambridge Tracts in Math. and Math. Physics. Cambridge Univ. Press.

[MacTutor, 2003] MacTutor (2003). The MacTutor History of Mathematics archive. `http://www-groups.dcs.st-and.ac.uk/~history`. University of St. Andrews.

[Mairhuber, 1956] Mairhuber, J. C. (1956). On Haar's theorem concerning Chebychev approximation problems having unique solutions. *Proc. Am. Math. Soc.*, 7:609–615.

[Marcus and Minc, 1969] Marcus, M. and Minc, H. (1969). *A Survey of Matrix Theory and Matrix Inequalities*. Prindle, Weber & Schmidt. Paperback reprint, Dover Publications, 1992.

[Micchelli, 1995] Micchelli, C. A. (1995). *Mathematical Aspects of Geometric Modeling*, volume 65 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM.

[Möller, 1988] Möller, H. M. (1988). On the construction of Gröbner bases using syzygies. *J. Symbolic Comput.*, 6:345–359.

[Möller and Sauer, 2000a] Möller, H. M. and Sauer, T. (2000a). H–bases for polynomial interpolation and system solving. *Advances Comput. Math.*, 12(4):335–362. to appear.

[Möller and Sauer, 2000b] Möller, H. M. and Sauer, T. (2000b). H–bases I: The foundation. In Cohen, A., Rabut, C., and Schumaker, L. L., editors, *Curve and Surface fitting: Saint–Malo 1999*, pages 325–332. Vanderbilt University Press.

[Möller and Sauer, 2004] Möller, H. M. and Sauer, T. (2004). Multivariate refinable functions of high approximation order via quotient ideals of Laurent polynomials. *Adv. Comput. Math.*, 20:205–228.

[Möller and Tenberg, 2001] Möller, H. M. and Tenberg, R. (2001). Multivariate polynomial system solving using intersections of eigenspaces. *J. Symbolic Comput.*, 32:513–531.

[Mourrain, 2016] Mourrain, B. (2016). Polynomial-exponential decomposition from moments. arXiv:1609.05720v1.

[Newton, 1687] Newton, I. (1687). *Philosphiae Naturalis Principia Mathematica*. Jussu Societatis Regiae ac typis Josephi Streater.

[Park and Woodburn, 1995] Park, H. and Woodburn, C. (1995). An algorithmic proof of Suslin's stability theorem for polynomial rings. *J. Algebra*, 178:277–298.

[Park, 1995] Park, H.-J. (1995). *A Computational Theory of Laurent Polynomial Rings and Multidimensional FIR Systems*. PhD thesis, University of California at Berkeley.

[Peters and Reif, 2008] Peters, J. and Reif, U. (2008). *Subdivision Surfaces*. Geometry and Computing. Springer.

[Plonka and Tasche, 2014] Plonka, G. and Tasche, M. (2014). Prony methods for recovery of structured functions. *GAMM–Mitt.*, 37:239–258.

[Potts and Tasche, 2015] Potts, D. and Tasche, M. (2015). Fast ESPRIT algorithms based on partial singular value decompositions. *Appl. Numer. Math.*, 88:31–45.

[Prony, 1795] Prony, C. (1795). Essai expérimental et analytique sur les lois de la dilabilité des fluides élastiques, et sur celles de la force expansive de la vapeur de l'eau et de la vapeur de l'alkool, à différentes températures. *J. de l'École polytechnique*, 2:24–77.

[Radon, 1948] Radon, J. (1948). Zur mechanischen Kubatur. *Monatsh. Math.*, 52:286–300.

[Roy and Kailath, 1989] Roy, R. and Kailath, T. (1989). ESPRIT – estimation of signal parameters via rotational invariance techniques. *IEEE Trans. Acoustics, Speech and Signal Processing*, 37:984–995.

[Sauer, 2001] Sauer, T. (2001). Gröbner bases, H–bases and interpolation. *Trans. Amer. Math. Soc.*, 353:2293–2308.

[Sauer, 2004] Sauer, T. (2004). Lagrange interpolation on subgrids of tensor product grids. *Math. Comp.*, 73:181–190.

[Sauer, 2006] Sauer, T. (2006). Polynomial interpolation in several variables: Lattices, differences, and ideals. In Buhmann, M., Hausmann, W., Jetter, K., Schaback, W., and Stöckler, J., editors, *Multivariate Approximation and Interpolation*, pages 189–228. Elsevier.

[Sauer, 2017] Sauer, T. (2017). Prony's method in several variables. *Numer. Math.*, 136:411–438. arXiv:1602.02352.

[Sauer, 2018a] Sauer, T. (2018a). Companion matrices and joint eigenvectors of commuting families of matrices for polynomial zero finding. *Monografías Matemáticas García de Galdeano*, (41):171—185.

[Sauer, 2018b] Sauer, T. (2018b). Prony's method in several variables: symbolic solutions by universal interpolation. *J. Symbolic Comput.*, 84:95–112. arXiv:1603.03944.

[Sauer, 2019] Sauer, T. (2019). Continued Fractions. Lecture notes, University of Passau.

[Sauer and Xu, 1996] Sauer, T. and Xu, Y. (1996). Regular points for Lagrange interpolation on the unit disk. *Numer. Algo.*, 12:287–296.

[Schmidt, 1986] Schmidt, R. (1986). Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34:276–280.

[Schmüdgen, 2017] Schmüdgen, K. (2017). *The Moment Problem*. Graduate Texts in Mathematics. Springer.

[Stetter, 2005] Stetter, H. J. (2005). *Numerical Polynomial Algebra*. SIAM.

[Trinks, 1978] Trinks, W. (1978). Über B. Buchbergers Verfahren, Systeme algebraischer Gleichungen zu lösen. *J. Number Theory*, 10:475–488.

[Vetterli and Kovačević, 1995] Vetterli, M. and Kovačević, J. (1995). *Wavelets and Subband Coding*. Prentice Hall.

[Warren, 2001] Warren, J. (2001). *Subdivision Methods for Geometric Design: A Constructive Approach*. Morgan Kaufmann Series in Computer Graphics and Geometric Modeling. Morgan Kaufman.

[Wilkinson, 1984] Wilkinson, J. H. (1984). The perfidious polynomial. In Golub, G. H., editor, *Studies in Numerical Analysis*, volume 24 of *MAA Studies in Mathematics*, pages 1–28. The Mathematical Association of America.

[Yosida, 1965] Yosida, K. (1965). *Functional Analysis*. Grundlehren der mathematischen Wissenschaften. Springer–Verlag.

# Index