Technische Universität Berlin

Fakultät II

Institut für Mathematik

Black box factorization of multivariate polynomials

Bachelorarbeit

zur Erlangung des Grades

Bachelor of Science

im Studiengang Mathematik

vorgelegt von

Sascha Timme

(Matrikelnummer 348922)

Berlin, August 2015

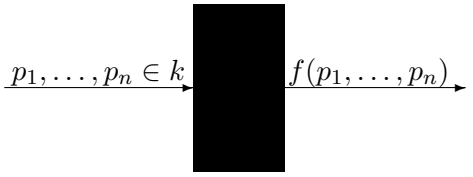
Erstgutachter: Prof. Dr. Peter Bürgisser
Zweitgutachter: Prof. Dr. Martin Skutella

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und eigenhändig sowie ohne unerlaubte fremde Hilfe und ausschließlich unter Verwendung der aufgeführten Quellen und Hilfsmittel angefertigt habe.
Die selbständige und eigenständige Anfertigung versichert an Eides statt:

Berlin, den 31. August 2015

Deutsche Zusammenfassung

Das Thema dieser Arbeit ist ein von Kaltofen und Trager [13] entwickelter Monte Carlo Algorithmus zur Faktorisierung multivariater Polyonome, die durch eine *black box* gegeben sind. Dabei ist die black box eines Polynoms $f \in k[X_1, \ldots, X_n]$ über einem Körper k ein Programm, welches als Eingabe $p_1, \ldots, p_n \in k$ hat und den Wert $f(p_1, \ldots, p_n)$ ausgibt:



Das bemerkenswerte an dem Algorithmus ist, dass die Laufzeit polynomiell in dem Grad des Eingabepolynoms, der Anzahl der black box Aufrufe und der Anzahl der Unbekannten ist. Zum Verständnis des Algorithmus werden in dieser Arbeit zunächst die wesentlichen theoretischen Grundlagen erarbeitet. Diese sind Hensel Lifting und eine effektive Version von Hilberts Irreduzibilitätstheorem.

Dabei ist Hensel Lifting ein Verfahren mit dem die Faktorisierung eines Polynoms über einem vollständigen lokalen noetherschen Ring aus der Faktorisierung im Quotientenring bezüglich des maximalen Ideals rekonstruiert werden kann. Dafür werden zunächst die Konzepte eines bewerteten Rings und der Vervollständigung eines Rings präsentiert. Ein besonderes Augenmerk liegt dabei auf der Vervollständigung von noetherschen Ringen bezüglich eines maximalen Ideals. Mit Hilfe dieser Konzepte wird dann rein algebraisch die Taylorentwicklung eines Polynoms hergeleitet und anschließend eine algebraische Version des (mehrdimensionalen) Newton-Verfahrens über bestimmten bewerteten Ringen entwickelt. Dieses hat wie die aus der Analysis bekannte Variante des Newtons-Verfahrens für einen geeigneten Startwert eine quadratische Konvergenz. Zusätzlich kann jedoch auch garantiert werden, dass die Jacobi Matrix in allen Iterationsschritten invertierbar bleibt. Anschließend wird das Konzept der Resultante und der Sylvester Matrix zweier Polynome präsentiert. Mit diesem wird dann das Hensel Lifting als ein Spezialfall des Newton-Verfahrens über vollständigen lokalen noetherschen Ringen hergeleitet und ein Hensel Lifting Algorithmus für noethersche Ringe entwickelt.

Zudem wird eine von Kaltofen entwickelte effektive Version von Hilberts Irreduzibilitätstheorem präsentiert. Mit Hilfe dieses Ergebnisses wird dann gezeigt, dass für ein multivariates Polynoms über einem perfekten Körper durch eine bestimmte

Substitution ein bivariates Polynom konstruiert werden kann, welches mit kontrollierbarer hoher Wahrscheinlichkeit das gleiche Faktorisierungsmuster wie das ursprüngliche multivariate Polynom aufweist.

Abschließend wird der detaillierte Faktorisierungsalgorithmus präsentiert und die Korrektheit, die Fehlschlagswahrscheinlichkeit und die polynomielle Laufzeit des Algorithmus bewiesen.

Contents

Chapter 1

Introduction

1.1 The problem of factoring polynomials

The problem of factoring polynomials is centuries old. In 1673 Newton already taught about computing factors of polynomials and this method was subsequently published in his Arithmetica Universalis [17]. In 1882 Kronecker [14] reduced the problem of factoring multivariate polynomials over finite extensions of the rational numbers (*algebraic number fields*) to factoring univariate polynomials over the integers, for which he then applied Newton's method. But implementations in early computer programs showed that these algorithms are not very practical for large problems which van der Waerden already discussed in his influential text Modern Algebra [18] in 1953. A theoretical and practical breakthrough was achieved by Elwyn Berlekamp during his time as a mathematical researcher at Bell Labs. He invented in 1967 [2] and improved in 1970 [3] an algorithm to factor univariate polynomials over finite fields. This algorithm was remarkable in several aspects. Firstly, it factors polynomials in time proportional to the cube of the input degree and was therefore the first algorithm which was suitable for use in applications. This also gave the first evidence that the problem of factoring polynomials is not as hard as the problem of factoring integers. In addition Berlekamp introduced the concept of *probabilistic algorithms*. He discovered that if one allows an algorithm to make random choices, e.g. pick a randomly chosen element out of a set, the algorithm could be sped up exponentially. The downside of this randomization is that the algorithm can fail or return a wrong result. If one can prove that the algorithm returns the correct output with a controllable high probability, it is called a *Monte Carlo algorithm*. In practice the performance of randomized algorithms to factor univariate polynomials over finite fields is far superior to any known deterministic algorithm.

The progress in factoring polynomials over finite fields suggested to apply these algorithms to the problem of factoring polynomials with integer coefficients. The idea is to factor the polynomial over a suitable finite field and then to reconstruct the integral factors from the modular images. One approach is to consider a finite field with a sufficiently large characteristic and another one is to make use of the Chinese remainder

theorem and to consider different modular images. Another approach was introduced by Zassenhaus [21] in 1969. He pointed to the p-adic numbers and "Hensel's Lemma", which were introduced by Hensel [8] in 1908. The described procedure is now called Hensel lifting and one of the standard techniques in computer algebra. But actually Gauß has preempted them all. In his Nachlass we can find an explicit description of a lifting procedure modulo prime powers [1], which is the core idea of Hensel's procedure. While the algorithm introduced by Zassenhaus has for most of the input polynomials a polynomially runtime, for some polynomials the algorithm has an exponential complexity due to "combinatorial explosion" in the lifting procedure. Nevertheless, the algorithm works well in practice and is implemented in many computer algebra systems. In 1982 A. Lenstra, H. Lenstra and L. Lovász [16] published a remarkable algorithm, called LLL algorithm, in which they solved the combinatorial explosion problem in the case of rational coefficients. This led to the development of algorithms to factor univariate polynomials over algebraic number fields in polynomial time [15].

1.2 The representation problem

In order to compute with polynomials we have to answer the question of how we uniquely represent a polynomial in our computer program. We call this the *data structure* or *representation* of a polynomial. For a polynomial the list of all terms with total degree less or equal than the degree of the polynomial is the *dense representation* of this polynomial. Consider the polynomial

$$f = X^3 + 2Y^2 - Z^2 \in \mathbb{Q}[X, Y, Z] \; . \tag{1.1}$$

It has the dense representation

$$\begin{aligned} f = {} & 1 \cdot X^3 + 0 \cdot X^2Y + 0 \cdot X^2Z + +0 \cdot X^2 + 0 \cdot XY^2 + 0 \cdot XYZ + 0 \cdot XY \\ & + 0 \cdot XZ^2 + 0 \cdot XZ + 0 \cdot X + 0 \cdot Y^3 + 0 \cdot Y^2Z + 2 \cdot Y^2 + 0 \cdot YZ^2 \\ & + 0 \cdot YZ + 0 \cdot Y + 0 \cdot Z^3 + (-1) \cdot Z^2 + 0 \cdot Z + 0 \cdot 1 \; . \end{aligned}$$

All algorithms so far assumed that the input polynomial has as a representation the dense representation. In 1985 Kaltofen [11] showed that the problem of factoring a multivariate polynomial can be reduced to the problem of factoring an univariate polynomial in polynomial time in the length of the dense representation. Therefore we can factor a multivariate polynomial over an algebraic number field in polynomial time in the length of its dense representation. But is this a satisfying result? Even for our previous polynomial (1.1) the dense representation has already 20 entries. In fact a polynomial in n indeterminates and with total degree d has

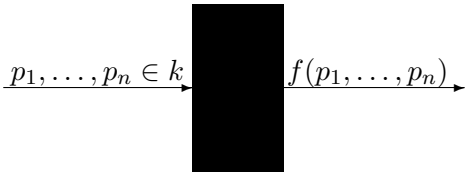
$$\sigma_{n,d} := \binom{n+d}{n}$$

[1] more details can be found in [19], p. 460

terms of total degree less than or equal d. Since $\sigma_{n,d}$ grows exponentially it follows that the length of the dense representation of a multivariate polynomial also grows exponentially! Thus the problem of factoring multivariate polynomials is obviously not satisfactorily solved.

Can we get a better result if we consider another representation? A more concise and readable representation is the *sparse representation* of a polynomial as in (1.1). It consists in general of a list of coefficients and exponents $(a_k, e_{k,1}, \ldots, e_{k,n})$ of the non-zero terms of the polynomial. We also have to consider the degree of the polynomial in the length of the sparse representation. Thus the convention is that the length of a list entry $(a_k, e_{k,1}, \ldots, e_{k,n})$ is the sum of the lengths of the exponents and the length of a_k. While this representation is elegant and the natural mathematical notation; it is unfortunately less suitable for computation.

A powerful technique to overcome this hurdle is to consider another representation called the *black box representation*. The black box representation $\mathcal{B}_f$ of a polynomial $f \in k[X_1, \ldots, X_n]$ is a program which accepts inputs $p_1, \ldots, p_n \in k$ and returns the value $f(p_1, \ldots, p_n)$:



The black box representation has several advantages. At first it is easy to construct from a sparse representation of a polynomial the corresponding black box. With sparse interpolation techniques, e.g. [1], it is also possible to get the sparse representation of a polynomial from a given black box in polynomial time. Furthermore, there are no constraints to the computation of the return value. For example there can be advantages if the polynomial is the determinant of a matrix. With the black box representation it would then be possible to use fast determinant algorithms to compute the return value. Moreover, it is possible that the black box representation uses even less memory space than the corresponding sparse representation.

With the framework of a black box representation it is now possible to solve several problems. This includes the factorization and gcd of multivariate polynomials in random polynomial time in the length of the total degree, number of variables and number of calls to the black box. Efficient Monte Carlo algorithms for some problems (including factorization and gcd) were introduced in an remarkable paper by Kaltofen und Trager [13] in 1990. The black box factorization algorithm proposed in this paper, and the necessary theoretical background, is the topic of this thesis.

1.3 Black box factorization

We want to have a first glimpse on the factorization algorithm to motivative the following theoretical chapters.

Assume we have a multivariate polynomial $f \in k[X_1, \ldots, X_n]$ over a field k with characteristic 0 and write $h_1^{e_1} \cdots h_r^{e_r}$ for the factorization of f. The factorization algorithm has as its input the black box $\mathcal{B}_f$ and it returns the multiplicities $e_1, \ldots, e_r$ and the following program.

For input $p_1, \ldots, p_n \in k$ it returns the values $h_1(p_1, \ldots, p_n), \ldots, h_r(p_1, \ldots, p_n)$.

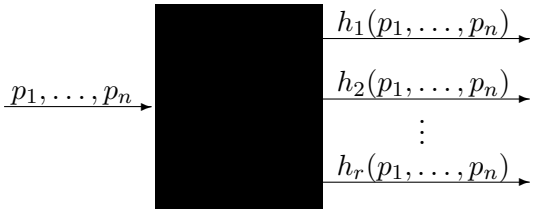


FIGURE 1.1: Output program

Furthermore, we assume that we can efficiently factor polynomials in $k[X_1, X_2]$.

The algorithm combines two ideas. The first idea is to use an effective version of the Hilbert irreducibility theorem which was first stated and proved by Hilbert [9] in 1892. The theorem states that for an irreducible polynomial $g(X, Y) \in \mathbb{Q}[X, Y]$, for almost all $a \in \mathbb{Q}$, the polynomial $g(a, Y) \in \mathbb{Q}[Y]$ is also irreducible. This can be generalized for irreducible multivariate polynomials $g \in \mathbb{Q}[X_1, \ldots, X_n, Y]$ such that for almost all $a_1, \ldots, a_n \in \mathbb{Q}$ the evaluation $g(a_1, \ldots, a_n, Y)$ remains irreducible in $\mathbb{Q}[Y]$. We need to quantify the "for almost all' part, i.e., an effective version, of the theorem and we have to be able to apply the theorem not only to polynomials over $\mathbb{Q}$. Unfortunately there is no known effective univariate version and the statement is clearly not applicable for important fields like finite fields or the complex numbers. But the situation can be rescued. In 1985 Kaltofen [10] constructed a substitution such that for an irreducible multivariate polynomial over a perfect field k the resulting bivariate polynomial remains irreducible with a controllable high probability. As a consequence for a multivariate polynomial $f \in k[X_1, \ldots, X_n]$ we can create a bivariate polynomial $f_2 \in k[X_1, X_2]$ such that each irreducible factor h_i of f has with a high probability a corresponding irreducible factor $g_{2,i}$ of f_2.

The second idea could be interpreted as an ansatz in homotopy continuation methods. We can construct a bivariate polynomial $\bar{f} \in k[X_1, Y]$ such that for $p_1, \ldots, p_n \in k$ and a known $\alpha \in k$ we have $\bar{f}(\alpha, 1) = f(p_1, \ldots, p_n)$ and $\bar{f}(X_1, 0) = f_2(X_1, 0)$. Since we can efficiently compute the factors $g_{2,i}$ of f_2 we have a factorization of $f_2(X_1, 0)$. With this factorization we can reconstruct the factors $\bar{g}_i$ of $\bar{f}$ with the by Zassenhaus introduced Hensel lifting. By construction we have then $\bar{g}_i(\alpha, 1) = h_i(p_1, \ldots, p_n)$.

This causes the following structure of this thesis. At first we derive the Hensel lifting algorithm in chapter 2 and then an effective version of Hilbert's irreducibility theorem in chapter 3. Finally we formulate the detailed black box factorization algorithm in chapter 4.

Chapter 2

Hensel lifting

Assume we want to factor the bivariate polynomial

$$f(X,Y) = Y^3 + (X-1)Y^2 + (-X+1)Y - 1 \in \mathbb{Q}[X][Y] \; .$$

The homotopy continuation ansatz is to transform this problem into a simpler one which we can easily solve, i.e., an univariate factorization problem. Hence consider

$$f(0,Y) = Y^3 - Y^2 + Y - 1 \in \mathbb{Q}[Y]$$

and in fact we can see that $f(0,Y) = (Y^2+1)(Y-1)$. Thus we are looking for polynomials

$$g(X,Y) = Y^2 + g_1(X)Y + g_0(X) \quad \text{and} \quad h(X,Y) = Y + h_0(X)$$

such that $f(X,Y) = g(X,Y)h(X,Y)$, $g(0,Y) = Y^2+1$ and $h(0,Y) = Y-1$. This is the same as solving the (non-linear) system of polynomial coefficient equations

$$\begin{aligned} 1 &= 1 \cdot 1 && (Y^3) \\ X-1 &= g_1 + h_0 && (Y^2) \\ -X+1 &= g_1 h_0 + g_0 && (Y) \\ -1 &= g_0 h_0 && (1) \end{aligned}$$

in $\mathbb{Q}[X]$.

A well-known method in numerical analysis to solve a system of non-linear functions is the Newton iteration (or Newton's method). It states

> Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be a differentiable function and $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$ a suitable start approximation such that $\|F(\boldsymbol{x}^{(0)})\| \leq \epsilon < 1$ and $J_F(\boldsymbol{x}^{(0)})$ invertible. Define the iteration
>
> $$\boldsymbol{x}^{(k+1)} := \boldsymbol{x}^{(k)} - J_F\left(\boldsymbol{x}^{(k)}\right)^{-1} F\left(\boldsymbol{x}^{(k)}\right)$$
>
> (if well defined) then $\|F(\boldsymbol{x}^{(k)})\| \leq \epsilon^{2^k}$ and $\|F(\lim_{k\to\infty} \boldsymbol{x}^{(k)})\| = 0$

If we were now able to give a corresponding method in our algebraic setting we could solve our system of coefficient equations and in fact this is possible. In order to do this we have to answer the following questions:

1. Can we define a notion of "closeness" or even a metric space over rings / modules?
2. Can we even define a notion of convergence? Do we also have a quadratic convergence?
3. Can we replace the analytical derivation?
4. Can we formulate an algebraic version of Taylor's formula?
5. What is a suitable initial approximate solution?
6. Can we guarantee that the Jacobian remains invertible?
7. Is the result also unique?

We will derive the Hensel lifting theorem for polynomials over an arbitrary Noetherian ring, although we only apply it in the concrete case that we have multivariate polynomials over a field of characteristic 0. I chose this approach because I derived on my own an algebraic version of the multivariate Newton iteration and the Hensel lifting theorem and it felt "natural" to make this in a more general setting. A drawback is that I needed some advanced results from commutative algebra in section 2.1 so that this thesis is not self-contained. But we will, as often as possible, refer to the concrete case of polynomial rings.

2.1 Valuation on a ring

First, all rings in this thesis are assumed to be commutative with identity 1.

We start with the fundamental question when two elements are "close". If not other mentioned we follow Bourbaki [5] in this section.

Definition 2.1 (Valuation)**.** Let A be a ring and Γ a totally ordered abelian group written multiplicatively. A *valuation* is a surjective map $\nu : A \mapsto \Gamma \cup \{0\} =: \Gamma_0$ which satisfies for all $a, b \in A$:

(1) $\nu(ab) = \nu(a)\nu(b)$ (multiplicative)

(2) $\nu(a+b) \leq \max\{\nu(a), \nu(b)\}$ (ultra-metric inequality)

(3) $\nu(1) = 1$ and $\nu(a) = 0$ if and only if $a = 0$

A valuation is called *discrete* if Γ is isomorph to $\mathbb{Z}$.

Remark 2.2. $\mathbb{Q}_{>0}$ with the usual multiplication is an example for Γ.

Remark 2.3. From the definition it follows that A has to be an integral domain.

Remark 2.4. It is possible to give a equivalent definition (and more common definition) for a valuation with a totally ordered *additive* abelian group Γ adjoined with ∞. In this case we have for all $a, b \in A$:

(1) $\nu(ab) = \nu(a) + \nu(b)$ (additive)

(2) $\nu(a+b) \geq \inf\{\nu(a), \nu(b)\}$

(3) $\nu(1) = 0$ and $\nu(a) = \infty$ if and only if $a = 0$

Remark 2.5. If $a \in A$ such that $a^n = 1$ for some integer $n \geq 1$, we have $\nu(a^n) = \nu(a)^n = 1$ by (1) and since Γ is a totally ordered multiplicative group we have $\nu(a) = 1$ for *every* valuation ν on A. In particular $\nu(-1) = 1$ and thus $\nu(-a) = \nu(a)$ for all $a \in A$. Since for $a \in A$ we have $0 = \nu(0) = \nu(a + (-a)) \leq \max\{\nu(a), \nu(-a)\} = \max\{\nu(a), \nu(a)\}$ by (2) it follows $\nu(a) \geq 0$ for all $a \in A$. If $a \in A$ is not zero then $\nu(a)\nu(a^{-1}) = \nu(aa^{-1}) = \nu(1) = 1$ and thus $\nu(a^{-1}) = 1/\nu(a)$.

Now we consider an important valuation, the $\mathfrak{m}$-adic valuation. Let A be a ring, $\mathfrak{m} \subset A$ a proper ideal. The sequence $(\mathfrak{m}^n)_{n\geq 0}$ of additive subgroups of A is called the $\mathfrak{m}$-*adic filtration* of A. Then the *order function* $\omega : A \to \mathbb{N} \cup \{\infty\},\ a \mapsto \omega(a)$ with

$$\omega(a) = n \ \Leftrightarrow a \in \mathfrak{m}^n \text{ and } a \notin \mathfrak{m}^{n+1} \tag{2.1}$$

$$\omega(a) = \infty \Leftrightarrow a \in \bigcap_{n\geq 0} \mathfrak{m}^n \tag{2.2}$$

is well defined.

The fact that the $\mathfrak{m}^n$ are additive subgroups of A implies that for $a, b \in A$

$$\omega(a+b) \geq \inf\{\omega(a), \omega(b)\}\ . \tag{2.3}$$

Since A is an integral domain it follows for $a, b \neq 0$ with $a \in \mathfrak{m}^r$, $a \notin \mathfrak{m}^{r+1}$ and $b \in \mathfrak{m}^s$, $b \notin \mathfrak{m}^{s+1}$ that $ab \notin \mathfrak{m}^{r+s+1}$ but $ab \in \mathfrak{m}^{r+s}$ and thus

$$\omega(a+b) = \omega(a) + \omega(b)\ . \tag{2.4}$$

Consider now the map

$$\nu : A \to \{2^{-k} \,|\, k \in \mathbb{N}\} \cup \{0\},\ a \mapsto 2^{-\omega(a)}\ ,$$

then it follows from the equations (2.1) - (2.4) that, if $\bigcap_{n\geq 0} \mathfrak{m}^n = \{0\}$, ν is a valuation map on A, called the $\mathfrak{m}$-*adic valuation* of A. In the concrete case of multivariate polynomials over a field this is clearly satisfied for every $\mathfrak{m}$.

A theorem from Krull [6] states that the ideal $\bigcap_{n\geq 0} \mathfrak{m}^n$ is $\{0\}$ if A is a Noetherian ring and no element of $1 + \mathfrak{m}$ is a divisor of 0 in A. In particular this is satisfied by Noetherian local rings.

Example 2.6. Let $A = \mathbb{Q}[X]$ and $\mathfrak{m} = (X)$. Then the order function is

$$\omega(a) = n \Leftrightarrow a \equiv 0 \bmod X^n \text{ and } a \not\equiv 0 \bmod X^{n+1}$$
$$\omega(0) = 0$$

With ν as the $\mathfrak{m}$-adic valuation we have

$$\nu(X-1)=1, \quad \nu(X)=\frac{1}{2} \quad \text{and} \quad \nu(X^6-X^3)=\frac{1}{8}\,.$$

Lemma 2.7 ([6])**.** Let A be a ring, Γ a totally ordered abelian group written multiplicatively and $\nu : A \mapsto \Gamma_0$ a valuation map. Then for $a_1, \ldots, a_n \in A$

$$\nu\Big(\sum_{j=1}^{n} a_j\Big) \leq \max_{1\leq j\leq n} \nu(a_j)\,. \tag{2.5}$$

Moreover, equality holds if there exists only a single index i such that $\nu(a_i) = \max_{1\leq j\leq n} \nu(a_j)$.
In particular we have for $a, b \in A$ with $\nu(a) \neq \nu(b)$ that $\nu(a+b) = \max\{\nu(a), \nu(b)\}$.

Proof. The inequality (2.5) follows with axiom (2) easily by induction over n. Now if there exists only a single index i such that $\nu(a_i) = \max_{1\leq j\leq n} \nu(a_j)$ then it follows with $x := \sum_{j\neq i} a_i$ and $y := \sum_{j=1}^{n} a_j$ by (2.5) that $\nu(x) < \nu(a_i)$ and $\nu(y) \leq \nu(a_i)$. Assume $\nu(y) < \nu(a_i)$. Since $a_i = y - x$ it follows that $\nu(a_i) \leq \max\{\nu(y), \nu(x)\} < \nu(a_i)$. This is clearly a contradiction. Hence $\nu(y) = \nu(a_i)$. □

Now consider again general valuations $\nu : A \mapsto \Gamma_0$ on a ring A. For $\boldsymbol{a} = (a_i) \in A^n$ define

$$\|\boldsymbol{a}\|_\nu := \max_{1\leq i\leq n} \nu(a_i)\,. \tag{2.6}$$

Then

$$d : A^n \times A^n \to \Gamma_0, \quad (a, b) \mapsto \|\boldsymbol{a}-\boldsymbol{b}\|_\nu \tag{2.7}$$

is an ultra-metric on A^n.

Proof. For $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \in A^n$ we have $d(a,b) = \|\boldsymbol{a}-\boldsymbol{b}\|_\nu = \max_i \nu(a_i - b_i) \geq 0$ and $d(a,b) = 0$ if and only if $\boldsymbol{a}-\boldsymbol{b} = \boldsymbol{0}$. Obviously $d(a,b) = d(b,a)$ and

$$\begin{aligned} d(\boldsymbol{a},\boldsymbol{c}) &= \|\boldsymbol{a}-\boldsymbol{c}\|_\nu = \|\boldsymbol{a}-\boldsymbol{b}+\boldsymbol{b}-\boldsymbol{c}\|_\nu \\ &= \max_i \nu(a_i - b_i + b_i - c_i) \\ &\leq \max_i \max\{\nu(a_i-b_i), \nu(b_i-c_i)\} \\ &= \max\{\|\boldsymbol{a}-\boldsymbol{b}\|_\nu, \|\boldsymbol{b}-\boldsymbol{c}\|_\nu\} = \max\{d(\boldsymbol{a},\boldsymbol{b}), d(\boldsymbol{b},\boldsymbol{c})\} \end{aligned}$$

□

If we identify $A^{m,n}$ with A^{mn} then (2.7) is also a metric on $A^{m,n}$. For $C = (c_{i,j}) \in A^{m,n}$ we set $\|C\|_\nu := \max_{\substack{1\leq i\leq m \\ 1\leq j\leq n}} \nu(c_{i,j})$.

Therefore every valuation ν induces a metric and thus a topology on the A-module A^n, $n \in \mathbb{N}_{\geq 1}$. If ν is the $\mathfrak{m}$-adic valuation on A then the topology on A^n, induced by (2.7), is called the $\mathfrak{m}$-*adic topology* and we write $\|\cdot\|_\mathfrak{m}$ instead of $\|\cdot\|_\nu$.

Now we introduce the concept of the completion of a ring which will prove to be very useful. Following Eisenbud [7] the *completion* $\hat{A}_\mathfrak{m}$ *of* A *with respect to the* $\mathfrak{m}$-*adic*

filtration is the *inverse limit* of the factor groups $A/\mathfrak{m}^i$. This is by definition a subgroup of the direct product

$$\hat{A}_\mathfrak{m} := \varprojlim A/\mathfrak{m}^i = \{a = (a_1, a_2, \dots) \in \prod_{i\in\mathbb{N}} A/\mathfrak{m}^i \mid a_j \equiv a_i \mod \mathfrak{m}^i \text{ for all } j > i\} .$$

Since each of the $A/\mathfrak{m}^i$ is a ring, $\hat{A}_\mathfrak{m}$ is also a ring. $\hat{A}_\mathfrak{m}$ has a filtration by ideals

$$\hat{\mathfrak{m}}_i := \{a = (a_1, a_2, \dots) \in \hat{A}_\mathfrak{m} \mid a_j = 0 \text{ for all } j \le i\}$$

and from the definitions it follows immediately that $\hat{A}_\mathfrak{m}/\hat{\mathfrak{m}}_i \cong A/\mathfrak{m}^i$. If the canonical inclusion map $A \to \hat{A}_\mathfrak{m},\ a \mapsto (a + \mathfrak{m}, a + \mathfrak{m}^2, \dots)$ is an isomorphism we shall say that A is *complete with respect to* $\mathfrak{m}$.

Proposition 2.8. Let $A = k[X_1, \dots, X_n]$ be the polynomial ring in n indeterminates over a field k and $\mathfrak{m} = (X_1, \dots, X_n)$ an ideal of A. Then the completion of A with respect to $\mathfrak{m}$ satisfies

$$\hat{A}_\mathfrak{m} \cong k[[X_1, \dots, X_n]],$$

where $k[[X_1, \dots, X_n]]$ denotes the ring of formal power series in n indeterminates.

Proof. With the maps

$$\varphi_i : k[[X_1, \dots, X_n]] \to k[X_1, \dots, X_n]/\mathfrak{m}^i\ ,\ f \mapsto f + \mathfrak{m}^i$$

we get the canonical map

$$\varphi : k[[X_1, \dots, X_n]] \to \hat{A}_\mathfrak{m} \subset \prod_i k[X_1, \dots, X_n]/\mathfrak{m}^i,$$
$$f \mapsto (\varphi_1(f), \varphi_2(f), \dots) = (f + \mathfrak{m}, f + \mathfrak{m}^2, \dots).$$

On the other hand we have that for each $(f_1 + \mathfrak{m}, f_2 + \mathfrak{m}^2, \dots) \in \hat{A}_\mathfrak{m}$ for all $j > i$

$$f_j = f_i + (\text{terms of degree} > i)\ .$$

Thus the map

$$\psi : \hat{A}_\mathfrak{m} \to k[[X_1, \dots, X_n]], \quad (f_1 + \mathfrak{m}, f_2 + \mathfrak{m}^2, \dots) \mapsto f_1 + (f_2 - f_1) + (f_3 - f_2) + \dots$$

is well defined and one checks immediately that $\varphi \circ \psi = id$ and $\psi \circ \varphi = id$. □

Remark 2.9. Furthermore, $\hat{\mathfrak{m}}_1 = \mathfrak{m} = (X_1, \dots, X_n)$ since $k[[X]]/\hat{\mathfrak{m}} \cong k[X]/\mathfrak{m} \cong k$.

Remark 2.10. $k[[X_1, \dots, X_n]]$ is also a unique factorization domain.

A useful property of the completion is that it inherits the Noetherian property of A.

Lemma 2.11. Let A be a Noetherian ring and $\mathfrak{m}$ an ideal of A. Then the completion $\hat{A}_\mathfrak{m}$ of A with respect to the $\mathfrak{m}$-adic filtration is also a Noetherian ring.

Proof. [6] □

Thus if A is a Noetherian ring then $\hat{\mathfrak{m}}_n = \hat{\mathfrak{m}}_1^n$ for all $n \in \mathbb{N}$ and we will briefly write $\hat{\mathfrak{m}}$ instead of $\hat{\mathfrak{m}}_1$ and $\hat{\mathfrak{m}}^n$ instead of $\hat{\mathfrak{m}}_n$.

Assume now that A is a Noetherian ring with ideal $\mathfrak{m}$. We want to determine when the $\hat{\mathfrak{m}}$-adic valuation on the completion $\hat{A}_\mathfrak{m}$ is well defined, i.e. $\bigcap_{n\geq 0}\mathfrak{m}^n = \{0\}$. By Krull's theorem this condition is satisfied if $\hat{A}_\mathfrak{m}$ is a local ring. Remember that a characterization of a local ring B is that for every element $b \in B$ is b or $1-b$ a unit.

Lemma 2.12. Let A be a ring, $\mathfrak{m} \subset A$ an ideal and $a \in A$. For positive integers i and j

$$a \text{ unit in } A/\mathfrak{m}^i \iff a \text{ unit in } A/\mathfrak{m}^j \ .$$

Proof. Let a be a unit in $A/\mathfrak{m}^i$ and first assume $i \geq j$. Then there exists a $b \in A$ such that $ab \equiv 1 \bmod \mathfrak{m}^i$ and since $\mathfrak{m}^j \subset \mathfrak{m}^i$ we have $ab \equiv 1 \bmod \mathfrak{m}^j$.

Now assume $i < j$ and without loss of generality $i = 1$ and $j = 2^k$ for some integer k. Then there exists a $b_0 \in A$ such that $ab_0 \equiv 1 \bmod \mathfrak{m}$. Now we can recursively define a sequence $(b_l) \subset A$ such that for $l \geq 1$

$$b_l \equiv 2b_{l-1} - ab_{l-1}^2 \bmod \mathfrak{m}^{2^l} \ . \tag{2.8}$$

We claim that then $ab_l \equiv 1 \bmod \mathfrak{m}^{2^l}$ for all $l \geq 0$. We prove the claim by induction over l. The induction start is already done and for the induction step $(l-1 \to l)$ consider

$$1 - ab_l \overset{(2.8)}{\equiv} 1 - a(2b_{l-1} - ab_{l-1}^2) \equiv 1 - 2ab_{l-1} + a^2b_{l-1}^2 \equiv (1-ab_{l-1})^2 \equiv 0 \bmod \mathfrak{m}^{2^l} \ .$$

Hence $ab_k \equiv 1 \bmod \mathfrak{m}^j$. □

Remark 2.13. Equation (2.8) gives an algorithm to efficiently compute the inverse of an element.

Lemma 2.14. Let A be a ring with maximal ideal $\mathfrak{m}$ and $\hat{A}_\mathfrak{m}$ its completion. Then $a = (a_1, a_2, \dots) \in \hat{A}_\mathfrak{m}$ is a unit if and only if $a_1 \neq 0$.

Proof. If $a_1 \neq 0$ each a_i is a unit in $A/\mathfrak{m}$ and therefore a unit in $A/\mathfrak{m}^i$ by lemma 2.12. Now it follows from $a_j \equiv a_i \bmod \mathfrak{m}^i$ for all $j > i$ that $a_j^{-1} \equiv a_i^{-1} \bmod \mathfrak{m}^i$ for all $j > i$ and we conclude that $b := (a_1^{-1}, a_2^{-1}, \dots) \in \hat{A}_\mathfrak{m}$ is the inverse of a.

Now suppose that $a \in \hat{A}_\mathfrak{m}$ is a unit. Then there exists a $b \in \hat{A}_\mathfrak{m}$ such that $ab = 1$ and in particular $a_1b_1 = 1$ and thus $a_1 \neq 0$. □

Hence the completion of a ring with respect to a maximal ideal is a local ring and we can conclude

Lemma 2.15. Let A be a Noetherian ring with maximal ideal $\mathfrak{m}$. Then the completion $\hat{A}_\mathfrak{m}$ of A with respect to the $\mathfrak{m}$-adic filtration is a Noetherian local ring with maximal ideal $\hat{\mathfrak{m}}$.

Proof. Since $\mathfrak{m}$ is a maximal ideal $\hat{A}_\mathfrak{m}/\hat{\mathfrak{m}} \cong A/\mathfrak{m}$ is a field and hence $\hat{\mathfrak{m}}$ a maximal ideal. Moreover, if $a = (a_1, a_2, \dots) \in \hat{A}_\mathfrak{m}$ not in $\hat{\mathfrak{m}}$, $a_1 \neq 0$ and by lemma 2.14 a unit. This shows that $\hat{A}_\mathfrak{m}$ is a local ring and $\hat{A}_\mathfrak{m}$ is also Noetherian by lemma 2.11. □

Remark 2.16. In particular $k[[X_1, \dots, X_n]]$ is a complete Noetherian local ring by our lemma and proposition 2.8.

The completion $\hat{A}_\mathfrak{m}$ of a Noetherian ring A with respect to a maximal ideal $\mathfrak{m}$ has therefore the $\hat{\mathfrak{m}}$-adic topology. Now we show that this notion of completion coincides with the notion of a complete metric space in the sense that every Cauchy sequence in $\hat{A}_\mathfrak{m}$ converges in $\hat{A}_\mathfrak{m}$.

Since $\hat{A}_\mathfrak{m}$ is a metric space, a series $(a_1, a_2, \dots) \subset \hat{A}_\mathfrak{m}$ converges in the $\hat{\mathfrak{m}}$-adic topology to an element $a \in \hat{A}_\mathfrak{m}$ if for every integer n there is an integer $i(n)$ such that $\|a - a_{i(n)}\|_{\hat{\mathfrak{m}}} \leq 2^{-n}$. This is equivalent to that for every integer n there is an integer $i(n)$ such that $a - a_{i(n)} \in \hat{\mathfrak{m}}^n$. Let $(a_i) \subset \hat{A}_\mathfrak{m}$ be a Cauchy sequence in the $\hat{\mathfrak{m}}$-adic topology. This means that for every integer n there exists an integer N such that

$$a_i - a_j \in \hat{\mathfrak{m}}^n \text{ for all } i, j \geq N \ .$$

This implies that for every integer n there exists an integer N such that $a_i \equiv a_N \bmod \hat{\mathfrak{m}}$ for $i \geq N$ and it follows immediately immediately that every Cauchy sequence converges in $\hat{A}_\mathfrak{m}$. Thus $\hat{A}^n_\mathfrak{m}$ is also complete since every sequence $\hat{A}^n_\mathfrak{m}$ is Cauchy if and only if the coordinate sequences $(a^{(i)}_j)$ are Cauchy.

We have seen that every Noetherian local ring yields in a natural way a valuation, the $\mathfrak{m}$-adic valuation. But it is of course possible to define valuations on other rings as well. This leads to the following definition.

Definition 2.17 (Valuation ring)**.** Let A be an integral domain and k its field of fractions. A is a *valuation ring* (or valued ring) if there exists a totally ordered multiplicative abelian group Γ and a valuation $\nu : k \mapsto \Gamma_0$ such that $A = \{\nu(x) \leq 1 \,|\, x \in k\}$. By definition, a field is not a valuation ring.

If ν is a discrete valuation A is called a *discrete valuation ring.*

Remark 2.18. It's again possible to give an equivalent definition for a valuation with a totally ordered *additive* abelian group Γ adjoined with ∞. Then the valuation ring is defined as $A := \{\nu(x) \geq 0 \,|\, x \in k\}$.

Remark 2.19. Let A be an integral domain and $\nu : A \mapsto \Gamma_0$ a valuation map. Denote by k the field of fractions of A and let a/b be an element of k. Then ν can easily be extended to a valuation of k with $\nu(a/b) := \nu(a)/\nu(b)$ and A is a subring of the valuation ring $A_\nu = \{\nu(x) \leq 1 \,|\, x \in k\}$.

Example 2.20. Let $k[[X]]$ be the ring of formal power series in one indeterminate over a field k. From proposition 2.8 it follows that $k[[X]]$ is a complete Noetherian local ring with maximal ideal $\mathfrak{m} = (X)$. We claim that $k[[X]]$ with the $\mathfrak{m}$-adic valuation ν is a valuation ring.
The field of fractions of $k[[X]]$ is $k((X)) = \{f/g \mid f, g \in k[[X]] \text{ with } g \neq 0\}$ and we can extend the $\mathfrak{m}$-adic valuation to $k((X))$ with $\nu(f/g) := \nu(f)/\nu(g) = 2^{-(\omega(f)-\omega(g))}$. Then the corresponding valuation ring is

$$\{f/g \in k((X)) \,|\, \nu(f/g) \leq 1\} = \{f/g \in k((X)) \mid \omega(f) - \omega(g) \geq 0\} \ .$$

For $f/g \in k((X))$ with $\nu(f/g) \leq 1$ we can assume that $\gcd(f, g) = 1$. Hence the only case that, at the first sight, f/g is not in $k[[X]]$ is $\omega(f) \geq \omega(g) = 0$. But then g has a non-zero constant coefficient g_0 and by lemma 2.14 g is a unit in $k[[X]]$. Therefore $f/g = (fg^{-1})/(gg^{-1}) = fg^{-1} \in k[[X]]$ and $k[[X]]$ is a valuation ring.

A valuation ring has the following useful property.

Lemma 2.21. Let A be a valuation ring with valuation map ν. Then $a \in A$ is a unit if and only if $\nu(a) = 1$.

Proof. Let $a \in A$ be a unit. Then there exists $a^{-1} \in A$ and $\nu(a) = 1/\nu(a^{-1})$. Since $a \in A$, $\nu(a) \leq 1$ implies $\nu(a^{-1}) \geq 1$ and with $a^{-1} \in A$ it follows $\nu(a^{-1}) = 1$ and thus $\nu(a) = 1$.
Now let a be in A with $\nu(a) = 1$. Denote by k the field of fractions of A and interpret a as an element of k. Then there exists $a^{-1} \in k$ and $\nu(a^{-1}) = 1/\nu(a) = 1$. Therefore $a^{-1} \in A$ and a is a unit. □

But this statement is not only true for valuation rings but also for Noetherian local rings A with maximal ideal $\mathfrak{m}$ and the $\mathfrak{m}$-adic valuation. Consider $a \in A$. Then $\nu(a) = 1$ if and only if $a \notin \mathfrak{m}$ and a characterization of a local ring is that every element not in $\mathfrak{m}$ is a unit.

This property yields also to the following valuable statement.

Lemma 2.22. Let A be a ring with valuation map ν such that $a \in A$ is a unit if and only if $\nu(a) = 1$. Let $a, b \in A$ with $\nu(b - a) < 1$. Then a is a unit if and only if b is a unit.

Proof. It is sufficient to show that $\nu(a) = 1$ if and only if $\nu(b) = 1$. If $\nu(a) = 1$ then $\nu(b) = \nu(a+(b-a)) \overset{2.5}{=} \max\{\nu(a), \nu(b-a)\} = 1$ and by an analogous argument it follows that $\nu(a) = 1$ if $\nu(b) = 1$. □

2.2 Taylor's formula

As a next step we derive an algebraic version of Taylor's formula (following Bourbaki's [4] neat derivation) and introduce the concept of a formal derivative as an replacement for the analytical derivative. But at first we have to fix some multi-index notation. Let $n \in \mathbb{N}$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n) \in \mathbb{N}^n$ and $\boldsymbol{X} = (X_1, \ldots, X_n)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ two families of indeterminates. We set $\boldsymbol{X} + \boldsymbol{Y} := (X_1 + Y_1, \ldots, X_n + Y_n)$, $\boldsymbol{X}^{\boldsymbol{\alpha}} := {X_1}^{\alpha_1} \cdot \ldots \cdot {X_n}^{\alpha_n}$, $\boldsymbol{\alpha}! := \alpha_1!\alpha_2! \ldots \alpha_n!$, $|\boldsymbol{\alpha}| := \alpha_1 + \alpha_2 + \ldots + \alpha_n$ and $\binom{\boldsymbol{\alpha}}{\boldsymbol{\beta}} := \binom{\alpha_1}{\beta_1} \ldots \binom{\alpha_n}{\beta_n}$ where $\binom{\alpha_i}{\beta_i}$ denotes the binomial coefficient.

Lemma 2.23 (Multi-index binomial theorem)**.** Let $n \in \mathbb{N}$, $\boldsymbol{X} = (X_1, \ldots, X_n)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ two families of indeterminates and $\boldsymbol{\alpha} \in \mathbb{N}^n$. Then

$$(\boldsymbol{X} + \boldsymbol{Y})^{\boldsymbol{\alpha}} = \sum_{\boldsymbol{0} \leq \boldsymbol{\beta} \leq \boldsymbol{\alpha}} \binom{\boldsymbol{\alpha}}{\boldsymbol{\beta}} \boldsymbol{X}^{\boldsymbol{\beta}} \boldsymbol{Y}^{\boldsymbol{\alpha}-\boldsymbol{\beta}} .$$

Proof. We have

$$\begin{aligned}(\boldsymbol{X}+\boldsymbol{Y})^{\boldsymbol{\alpha}} &= \prod_{i=0}^{n}(X_i+Y_i)^{\alpha_i}\\ &= \prod_{i=0}^{n}\sum_{\beta_i=0}^{\alpha_i}\binom{\alpha_i}{\beta_i}X_i^{\beta_i}Y_i^{\alpha_i-\beta_i}\\ &= \Big(\sum_{\beta_1=0}^{\alpha_1}\binom{\alpha_1}{\beta_1}X_1^{\beta_1}Y_1^{\alpha_1-\beta_1}\Big)\cdots\Big(\sum_{\beta_n=0}^{\alpha_n}\binom{\alpha_n}{\beta_n}X_n^{\beta_n}Y_n^{\alpha_n-\beta_n}\Big)\,.\end{aligned}$$

Now we can expand and rearrange the product:

$$\begin{aligned}&= \sum_{\beta_1=0}^{\alpha_1}\cdots\sum_{\beta_n=0}^{\alpha_n}\binom{\alpha_1}{\beta_1}\ldots\binom{\alpha_n}{\beta_n}X_1^{\beta_1}\ldots X_n^{\beta_n}Y_1^{\alpha_1-\beta_1}\ldots Y_n^{\alpha_n-\beta_n}\\ &= \sum_{\mathbf{0}\le\boldsymbol{\beta}\le\boldsymbol{\alpha}}\binom{\boldsymbol{\alpha}}{\boldsymbol{\beta}}\boldsymbol{X}^{\boldsymbol{\beta}}\boldsymbol{Y}^{\boldsymbol{\alpha}-\boldsymbol{\beta}}\,.\end{aligned}$$

□

Definition 2.24. Let n be an integer, A an integral domain and $\boldsymbol{X}=(X_1,\ldots,X_n)$ and $\boldsymbol{Y}=(Y_1,\ldots,Y_n)$ two families of indeterminates. For $f\in A[\boldsymbol{X}]=A[X_1,\ldots,X_n]$ we can consider $f(\boldsymbol{X}+\boldsymbol{Y})$ as a polynomial in $A[\boldsymbol{X}][\boldsymbol{Y}]$ and denote for all $\boldsymbol{\alpha}\in\mathbb{N}^n$ by $\Delta^{\boldsymbol{\alpha}}f\in A[\boldsymbol{X}]$ the coefficient of $\boldsymbol{Y}^{\boldsymbol{\alpha}}$ in $f(\boldsymbol{X}+\boldsymbol{Y})$.

Remark 2.25. From the definition of $\Delta^{\boldsymbol{\alpha}}f$ it follows immediately that $\Delta^{\boldsymbol{\alpha}}\in\mathrm{End}(A[\boldsymbol{X}])$ where $A[\boldsymbol{X}]$ is considered as an A-module.

Example 2.26. Consider again our example

$$\begin{aligned}f(X,Y) &= Y^3+(X-1)Y^2+(-X+1)Y-1\\ &= Y^3+XY^2-Y^2-XY+Y-1\,.\end{aligned}$$

Then

$$\begin{aligned}f(X+Z_1,Y+Z_2) &= (Y+Z_2)^3-(Y+Z_2)^2+(X+Z_1)(Y+Z_2)^2+(Y+Z_2)\\ &\quad-(X+Z_1)(Y+Z_2)-1\\ &= Z_2^3+Z_2^2Z_1+(3Y+X-1)Z_2^2+(2Y-1)Z_1Z_2\\ &\quad+(3Y^2+2XY-2Y-X+1)Z_2+(Y^2-Y)Z_1\\ &\quad+Y^3+XY^2-Y^2-XY+Y-1\end{aligned}$$

Hence, $\Delta^{0,1}f(X,Y)=3Y^2+2XY-2Y-X+1$ and $\Delta^{1,1}f(\boldsymbol{X})=2Y-1$.

In the following all summations are about $\mathbb{N}^n$. Let $f\in A[\boldsymbol{X}]=A[X_1,\ldots,X_n]$ be a multivariate polynomial and by definition

$$f(\boldsymbol{X}+\boldsymbol{Y})=\sum_{\boldsymbol{\alpha}}\Delta^{\boldsymbol{\alpha}}f(\boldsymbol{X})\boldsymbol{Y}^{\boldsymbol{\alpha}}\,. \tag{2.9}$$

If we substitute $\boldsymbol{X}\mapsto\boldsymbol{a}$ and $\boldsymbol{Y}\mapsto\boldsymbol{X}-\boldsymbol{a}$ for $\boldsymbol{a}\in A^n$ we have

$$f(\boldsymbol{X}) = \sum_{\boldsymbol{\alpha}} \Delta^{\boldsymbol{\alpha}} f(\boldsymbol{a})(\boldsymbol{X} - \boldsymbol{a})^{\boldsymbol{\alpha}} \,. \tag{2.10}$$

Since for $g \in A[\boldsymbol{X}]$

$$\begin{aligned}(fg)(\boldsymbol{X}+\boldsymbol{Y}) &= \left(\sum_{\boldsymbol{\alpha}} \Delta^{\boldsymbol{\alpha}} f(\boldsymbol{X})\boldsymbol{Y}^{\boldsymbol{\alpha}}\right)\left(\sum_{\boldsymbol{\beta}} \Delta^{\boldsymbol{\beta}} g(\boldsymbol{X})\boldsymbol{Y}^{\boldsymbol{\beta}}\right) \\ &= \sum_{\boldsymbol{\gamma}} \left[\sum_{\boldsymbol{\alpha}+\boldsymbol{\beta}=\boldsymbol{\gamma}} \Delta^{\boldsymbol{\alpha}} f(\boldsymbol{X})\Delta^{\boldsymbol{\beta}} g(\boldsymbol{X})\right] \boldsymbol{Y}^{\boldsymbol{\gamma}}\end{aligned}$$

we have

$$\Delta^{\boldsymbol{\gamma}}(fg)(\boldsymbol{X}) = \sum_{\boldsymbol{\alpha}+\boldsymbol{\beta}=\boldsymbol{\gamma}} \Delta^{\boldsymbol{\alpha}} f(\boldsymbol{X})\Delta^{\boldsymbol{\beta}} g(\boldsymbol{X}) \,. \tag{2.11}$$

Now let $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ be another family of indeterminates. Then we have

$$\begin{aligned}f(\boldsymbol{X}+\boldsymbol{Y}+\boldsymbol{Z}) &= f(\boldsymbol{X}+(\boldsymbol{Y}+\boldsymbol{Z})) \\ &= \sum_{\boldsymbol{\alpha}} \Delta^{\boldsymbol{\alpha}} f(\boldsymbol{X})(\boldsymbol{Y}+\boldsymbol{Z})^{\boldsymbol{\alpha}}\end{aligned} \tag{2.12}$$

and on the other hand

$$\begin{aligned}f(\boldsymbol{X}+\boldsymbol{Y}+\boldsymbol{Z}) &= \sum_{\boldsymbol{\beta}} \Delta^{\boldsymbol{\beta}} f(\boldsymbol{X}+\boldsymbol{Y})\boldsymbol{Z}^{\boldsymbol{\beta}} \\ &= \sum_{\boldsymbol{\beta}} \left[\sum_{\boldsymbol{\gamma}} \Delta^{\boldsymbol{\gamma}}(\Delta^{\boldsymbol{\beta}} f(\boldsymbol{X}))\boldsymbol{Y}^{\boldsymbol{\gamma}}\right] \boldsymbol{Z}^{\boldsymbol{\beta}} \\ &= \sum_{\boldsymbol{\beta},\boldsymbol{\gamma}} (\Delta^{\boldsymbol{\gamma}}\Delta^{\boldsymbol{\beta}} f)(\boldsymbol{X})\boldsymbol{Y}^{\boldsymbol{\gamma}}\boldsymbol{Z}^{\boldsymbol{\beta}} \,.\end{aligned} \tag{2.13}$$

By the multi-index binomial theorem 2.23 it follows

$$(\boldsymbol{Y}+\boldsymbol{Z})^{\boldsymbol{\alpha}} = \sum_{\boldsymbol{0}\le\boldsymbol{\beta}\le\boldsymbol{\alpha}} \binom{\boldsymbol{\alpha}}{\boldsymbol{\beta}} \boldsymbol{Y}^{\boldsymbol{\alpha}-\boldsymbol{\beta}}\boldsymbol{Z}^{\boldsymbol{\beta}} = \sum_{\boldsymbol{\gamma}+\boldsymbol{\beta}=\boldsymbol{\alpha}} \binom{\boldsymbol{\gamma}+\boldsymbol{\beta}}{\boldsymbol{\beta}} \boldsymbol{Y}^{\boldsymbol{\gamma}}\boldsymbol{Z}^{\boldsymbol{\beta}} = \sum_{\boldsymbol{\gamma}+\boldsymbol{\beta}=\boldsymbol{\alpha}} \frac{(\boldsymbol{\gamma}+\boldsymbol{\beta})!}{\boldsymbol{\gamma}!\boldsymbol{\beta}!} \boldsymbol{Y}^{\boldsymbol{\gamma}}\boldsymbol{Z}^{\boldsymbol{\beta}} \,.$$

Hence (2.12) becomes

$$\begin{aligned}\sum_{\boldsymbol{\alpha}} \Delta^{\boldsymbol{\alpha}} f(\boldsymbol{X})(\boldsymbol{Y}+\boldsymbol{Z})^{\boldsymbol{\alpha}} &= \sum_{\boldsymbol{\alpha}} \Delta^{\boldsymbol{\alpha}} f(\boldsymbol{X}) \sum_{\boldsymbol{\gamma}+\boldsymbol{\beta}=\boldsymbol{\alpha}} \frac{(\boldsymbol{\gamma}+\boldsymbol{\beta})!}{\boldsymbol{\gamma}!\boldsymbol{\beta}!} \boldsymbol{Y}^{\boldsymbol{\gamma}}\boldsymbol{Z}^{\boldsymbol{\beta}} \\ &= \sum_{\boldsymbol{\gamma},\boldsymbol{\beta}} \frac{(\boldsymbol{\gamma}+\boldsymbol{\beta})!}{\boldsymbol{\gamma}!\boldsymbol{\beta}!} \Delta^{\boldsymbol{\gamma}+\boldsymbol{\beta}} f(\boldsymbol{X})\boldsymbol{Y}^{\boldsymbol{\gamma}}\boldsymbol{Z}^{\boldsymbol{\beta}}\end{aligned}$$

and with (2.13) we get

$$\Delta^{\boldsymbol{\gamma}}\Delta^{\boldsymbol{\beta}} f = \frac{(\boldsymbol{\gamma}+\boldsymbol{\beta})!}{\boldsymbol{\gamma}!\boldsymbol{\beta}!} \Delta^{\boldsymbol{\gamma}+\boldsymbol{\beta}} f \,. \tag{2.14}$$

Before we proceed we have to introduce the concept of the formal derivative of a polynomial as a replacement of the analytical derivative.

Definition 2.27 (Formal derivative)**.** Let A be a (commutative) ring, n a positive integer and $\boldsymbol{X} = (X_1, \ldots, X_n)$ a family of indeterminates. For $\boldsymbol{\alpha} = (\alpha_i) \in \mathbb{N}^n$ the map

$$D_i : A[\boldsymbol{X}] \to A[\boldsymbol{X}], \quad \boldsymbol{X}^{\boldsymbol{\alpha}} \mapsto \begin{cases} \alpha_i X_i^{\alpha_i - 1} \prod\limits_{\substack{1 \le j \le n \\ i \ne j}} X_j^{\alpha_j} & , \alpha_i > 0 \\ 0 & , \alpha_i = 0 \end{cases} \tag{2.15}$$

is an A-linear ring homomorphism. For a polynomial $f \in A[\boldsymbol{X}]$ $D_i f$ is the *formal partial derivative* of f. It follows from (2.15) that $D_i D_j = D_j D_i$ for any $1 \le i, j \le n$. Thus for $\boldsymbol{\beta} = (\beta_i) \in \mathbb{N}^n$

$$D^{\boldsymbol{\beta}} : A[\boldsymbol{X}] \to A[\boldsymbol{X}], \quad \boldsymbol{X}^{\boldsymbol{\alpha}} \mapsto \begin{cases} \frac{\boldsymbol{\alpha}!}{(\boldsymbol{\alpha}-\boldsymbol{\beta})!} & , \ \boldsymbol{\alpha} \ge \boldsymbol{\beta} \\ 0 & , \text{ else} \end{cases} \tag{2.16}$$

is a well defined A-linear ring homomorphism. $D^{\boldsymbol{\beta}} f$ is called the *formal derivative* of f.

Remark 2.28. For a polynomial $f \in A[\boldsymbol{X}]$ the formal derivative coincides with the known analytical derivative. Moreover, computing rules like the product and chain rule also hold for the formal derivative.

Remark 2.29. We shall write, if it is more suitable, analog to the usual notation $\partial f / \partial X_i$ instead of $D_i f$.

Lemma 2.30. Let be n be a positive integer and $\boldsymbol{X} = (X_1, \ldots, X_n)$ a family of indeterminates. Then for all $f \in A[\boldsymbol{X}]$ and $\boldsymbol{\alpha} = (\alpha_i) \in \mathbb{N}^n$ it holds

$$D^{\boldsymbol{\alpha}} f = \boldsymbol{\alpha}! \Delta^{\boldsymbol{\alpha}} f \ .$$

Proof. The lemma is proven by induction over the length of $\boldsymbol{\alpha}$. Thus let $|\boldsymbol{\alpha}| = 1$. Then there exists an index $i \in \{1, \ldots, n\}$ such that $\boldsymbol{\alpha} = \boldsymbol{e}_i$ with $\alpha_i = 1$ and $\alpha_j = 0$ for all $j \ne i$. Now define for an arbitrary $\boldsymbol{\beta} \in \mathbb{N}^n$

$$p := X_i^{\beta_i} \quad \text{and} \quad q := \prod_{i \ne j} X_j^{\beta_j} \ .$$

By construction, $pq = \boldsymbol{X}^{\boldsymbol{\beta}}$ and with (2.11) we get

$$\begin{aligned} \Delta^{\boldsymbol{\alpha}} \boldsymbol{X}^{\boldsymbol{\beta}} &= \Delta^{\boldsymbol{e}_i} pq \\ &= \Delta^{\boldsymbol{e}_i} p \cdot \Delta^{\boldsymbol{0}} q + \Delta^{\boldsymbol{0}} p \cdot \Delta^{\boldsymbol{e}_i} q \ . \end{aligned} \tag{2.17}$$

Now we have

$$p(\boldsymbol{X} + \boldsymbol{Y}) = (X_i + Y_i)^{\beta_i} = \sum_{k=0}^{\beta_i} \binom{\beta_i}{k} X_i^{\beta_i - k} Y_i^k$$

and

$$q(\boldsymbol{X} + \boldsymbol{Y}) = \prod_{j \ne i} (X_j + Y_j)^{\beta_j} = \prod_{j \ne i} \sum_{k=0}^{\beta_j} \binom{\beta_j}{k} X_i^{\beta_j - k} Y_j^k \ .$$

Thus

$$\Delta^{\boldsymbol{e}_i} p = \begin{cases} \beta_i X_i^{\beta_i - 1} & \beta_i > 0 \\ 0 & \beta_i = 0 \end{cases} \quad \text{and} \quad \Delta^{\boldsymbol{0}} q = \prod_{j \ne i} (X_j)^{\beta_j} = q \ .$$

Since Y_i doesn't appear in $q(\boldsymbol{X}+\boldsymbol{Y})$ we have $\Delta^{\boldsymbol{e}_i}q = 0$. With (2.17) it follows that

$$\Delta^{\boldsymbol{\alpha}}\boldsymbol{X}^{\boldsymbol{\beta}} = \Delta^{e_i}\boldsymbol{X}^{\boldsymbol{\beta}} = \begin{Bmatrix} \beta_i X_i^{\beta_i-1}\prod_{j\neq i}(X_j)^{\beta_j} & , & \beta_i > 0 \\ 0 & , & \beta_i = 0 \end{Bmatrix} = D^{e_i}\boldsymbol{X}^{\boldsymbol{\beta}} = \Delta^{\boldsymbol{\alpha}}\boldsymbol{X}^{\boldsymbol{\beta}} .$$

Since $\Delta^{\boldsymbol{\alpha}} \in \mathrm{End}(A[\boldsymbol{X}])$ it follows that $\Delta^{\boldsymbol{\alpha}}f = D^{\boldsymbol{\alpha}}f = \boldsymbol{\alpha}!D^{\boldsymbol{\alpha}}f$.

Now let us assume that our induction hypothesis holds for $m \in \mathbb{N}$ and let $\boldsymbol{\alpha} \in \mathbb{N}^n$ with $|\boldsymbol{\alpha}| = m+1$. Notice that we obtain for $\boldsymbol{\beta}, \boldsymbol{\gamma} \in \mathbb{N}^n$ by (2.14)

$$(\boldsymbol{\gamma}!\Delta^{\boldsymbol{\gamma}})(\boldsymbol{\beta}!\Delta^{\boldsymbol{\beta}}) = (\boldsymbol{\gamma}+\boldsymbol{\beta})!\Delta^{\boldsymbol{\gamma}+\boldsymbol{\beta}} \in \mathrm{End}(A[\boldsymbol{X}]) . \tag{2.18}$$

Then there exists $i \in \{1,\dots,n\}$ such that $\boldsymbol{\alpha}-\boldsymbol{e}_i \in \mathbb{N}^n$ and by (2.18) we obtain

$$\boldsymbol{\alpha}!\Delta^{\boldsymbol{\alpha}} = ((\boldsymbol{\alpha}-\boldsymbol{e}_i)+\boldsymbol{e}_i)!\Delta^{(\boldsymbol{\alpha}-\boldsymbol{e}_i)+\boldsymbol{e}_i} = (\boldsymbol{\alpha}-\boldsymbol{e}_i)!\Delta^{(\boldsymbol{\alpha}-\boldsymbol{e}_i)} \circ \boldsymbol{e}_i!\Delta^{\boldsymbol{e}_i} .$$

Therefore it follows together with the induction hypothesis:

$$\begin{aligned} \boldsymbol{\alpha}!\Delta^{\boldsymbol{\alpha}}f &= ((\boldsymbol{\alpha}-\boldsymbol{e}_i)!\Delta^{(\boldsymbol{\alpha}-\boldsymbol{e}_i)} \circ \boldsymbol{e}_i!\Delta^{\boldsymbol{e}_i})f \\ &= (\boldsymbol{\alpha}-\boldsymbol{e}_i)!\Delta^{(\boldsymbol{\alpha}-\boldsymbol{e}_i)}(\boldsymbol{e}_i!\Delta^{\boldsymbol{e}_i}f) \\ &= (\boldsymbol{\alpha}-\boldsymbol{e}_i)!\Delta^{(\boldsymbol{\alpha}-\boldsymbol{e}_i)}(D^{\boldsymbol{e}_i}f) \\ &= D^{(\boldsymbol{\alpha}-\boldsymbol{e}_i)}(D^{\boldsymbol{e}_i}f) \\ &= D^{\boldsymbol{\alpha}}f \end{aligned}$$

□

Now everything is in place to formulate Taylor's formula in our algebraic setting.

Theorem 2.31 (Taylor's formula)**.** Let A be an integral domain, $n \in \mathbb{N}$, $\boldsymbol{X} = (X_1,\dots,X_n)$ a family of indeterminates and $f \in A[\boldsymbol{X}]$. Then we have for another family of indeterminates $\boldsymbol{Y} = (Y_1,\dots,Y_n)$

$$f(\boldsymbol{X}+\boldsymbol{Y}) = \sum_{\boldsymbol{\alpha}} \frac{1}{\boldsymbol{\alpha}!}(D^{\boldsymbol{\alpha}}f)(\boldsymbol{X})\boldsymbol{Y}^{\boldsymbol{\alpha}}$$

and for $\boldsymbol{a} \in A^n$

$$f(\boldsymbol{X}) = \sum_{\boldsymbol{\alpha}} \frac{1}{\boldsymbol{\alpha}!}(D^{\boldsymbol{\alpha}}f)(\boldsymbol{a})(\boldsymbol{X}-\boldsymbol{a})^{\boldsymbol{\alpha}} .$$

Proof. By (2.9) $f(\boldsymbol{X}+\boldsymbol{Y}) = \sum_{\boldsymbol{\alpha}} \Delta^{\boldsymbol{\alpha}}f(\boldsymbol{X})\boldsymbol{Y}^{\boldsymbol{\alpha}}$ and by lemma 2.30 it follows the first statement. By (2.10) $f(\boldsymbol{X}) = \sum_{\boldsymbol{\alpha}} \Delta^{\boldsymbol{\alpha}}f(\boldsymbol{a})(\boldsymbol{X}-\boldsymbol{a})^{\boldsymbol{\alpha}}$ and again by lemma 2.30 it follows the second statement. □

Finally, in preparation for the Newton iteration, we formulate a version of Taylor's formula for systems of polynomials.

Corollary 2.32. Let A be an integral domain, n a positive integer, $\boldsymbol{X} = (X_1,\dots,X_n)$ a family of indeterminates and $F(\boldsymbol{X}) = \begin{bmatrix} f_1(\boldsymbol{X}) \\ \vdots \\ f_n(\boldsymbol{X}) \end{bmatrix} \in A[\boldsymbol{X}]^n$ a system of polynomials. For

$\boldsymbol{a} \in A^n$ we have

$$F(\boldsymbol{X}) = F(\boldsymbol{a}) + J_F(\boldsymbol{a}) \cdot (\vec{\boldsymbol{X}} - \vec{\boldsymbol{a}}) + \sum_{|\boldsymbol{\alpha}|\geq 2} \frac{1}{\boldsymbol{\alpha}!} D^{\boldsymbol{\alpha}} F(\boldsymbol{a})(\boldsymbol{X} - \boldsymbol{a})^{\boldsymbol{\alpha}} \ .$$

where J_F is the Jacobian of F and $\vec{\boldsymbol{X}}$ and $\vec{\boldsymbol{a}}$ indicate that $\boldsymbol{X}$ and $\boldsymbol{a}$ should be interpreted as vectors.

Proof. For $\boldsymbol{a} \in A^n$ and $f_i \in A[\boldsymbol{X}]$ we have, by Taylor's formula 2.31,

$$\begin{aligned} f_i(\boldsymbol{X}) &= \sum_{\boldsymbol{\alpha}} \frac{1}{\boldsymbol{\alpha}} (D^{\boldsymbol{\alpha}} f_i)(\boldsymbol{a})(\boldsymbol{X} - \boldsymbol{a})^{\boldsymbol{\alpha}} \\ &= f_i(\boldsymbol{a}) + \sum_{|\boldsymbol{\alpha}|=1} (D^{\boldsymbol{\alpha}} f_i)(\boldsymbol{a})(\boldsymbol{X} - \boldsymbol{a})^{\boldsymbol{\alpha}} + \sum_{|\boldsymbol{\alpha}|\geq 2} \frac{1}{\boldsymbol{\alpha}} (D^{\boldsymbol{\alpha}} f_i)(\boldsymbol{a})(\boldsymbol{X} - \boldsymbol{a})^{\boldsymbol{\alpha}} \\ &= f_i(\boldsymbol{a}) + \sum_{j=1}^{n} (D_j f_i)(\boldsymbol{a})(X_j - a_j) + \sum_{|\boldsymbol{\alpha}|\geq 2} \frac{1}{\boldsymbol{\alpha}} (D^{\boldsymbol{\alpha}} f_i)(\boldsymbol{a})(\boldsymbol{X} - \boldsymbol{a})^{\boldsymbol{\alpha}} \ . \end{aligned}$$

Hence

$$\begin{aligned} F(\boldsymbol{X}) &= F(\boldsymbol{a}) + J_F(\boldsymbol{a}) \cdot (\vec{\boldsymbol{X}} - \vec{\boldsymbol{a}}) + \sum_{|\boldsymbol{\alpha}|\geq 2} \frac{1}{\boldsymbol{\alpha}!} \begin{bmatrix} D^{\boldsymbol{\alpha}} f_1(\boldsymbol{a})(\boldsymbol{X} - \boldsymbol{a})^{\boldsymbol{\alpha}} \\ \vdots \\ D^{\boldsymbol{\alpha}} f_n(\boldsymbol{a})(\boldsymbol{X} - \boldsymbol{a})^{\boldsymbol{\alpha}} \end{bmatrix} \\ &= F(\boldsymbol{a}) + J_F(\boldsymbol{a}) \cdot (\vec{\boldsymbol{X}} - \vec{\boldsymbol{a}}) + \sum_{|\boldsymbol{\alpha}|\geq 2} \frac{1}{\boldsymbol{\alpha}!} D^{\boldsymbol{\alpha}} F(\boldsymbol{a})(\boldsymbol{X} - \boldsymbol{a})^{\boldsymbol{\alpha}} \ . \end{aligned}$$

□

2.3 Newton iteration

Now that we have derived an algebraic version of Taylor's formula and introduced a metric space on rings with a valuation we are ready to state an algebraic version of the multi-dimensional Newton iteration. I developed this based on a one-dimensional version of the Newton Iteration in Modern Computer Algebra [20].

In this section let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a family of indeterminates and A a ring with valuation map ν such that for all $a \in A$ $\nu(a) \leq 1$ and that a is a unit if and only if $\nu(a) = 1$. This condition is satisfied for each valuation ring by lemma 2.21, but also for every Noetherian local ring A with maximal ideal $\mathfrak{m}$ and equipped with the $\mathfrak{m}$-adic valuation. Moreover, we abbreviate the elements $(a_1, \ldots, a_n) \in A^n$ as $\boldsymbol{a}$ and identify the A-module A^n and the cartesian product A^n with each other and use a suitable interpretation for $\boldsymbol{a}$. In particular for $f \in A[\boldsymbol{X}]$ we denote with $f(\boldsymbol{a})$ the evaluation of f at $(a_1, \ldots, a_n)$. Furthermore let $\|\cdot\|_\nu$ be the ultra-metric induced by ν which we defined in (2.6) and (2.7). Notice that $\nu(a)$ and $\|a\|_\nu$ coincide for $a \in A$.

At first we note the following useful estimate.

Lemma 2.33. For $C \in A^{m,n}$ and $\boldsymbol{a} \in A^n$ we have $\|C\boldsymbol{a}\|_\nu \leq \|C\|_\nu \|\boldsymbol{a}\|_\nu$

Proof. Let $C = (c_{i,j}) \in A^{n,n}$ be a matrix and $\boldsymbol{a} \in A^n$. Then

$$\begin{aligned}\|C\boldsymbol{a}\|_\nu &= \max_{1\leq i\leq m}\Big(\sum_{j=1} c_{i,j}a_j\Big) \leq \max_{1\leq i\leq m}\max_{\substack{1\leq i\leq m\\1\leq j\leq n}} \nu(c_{i,j}a_j)\\ &= \max_{\substack{1\leq i\leq m\\1\leq j\leq n}} \nu(c_{i,j})\nu(a_j) \leq \max_{\substack{1\leq i\leq m\\1\leq j\leq n}} \nu(c_{i,j}) \max_{1\leq j\leq n} \nu(a_j) = \|C\|_\nu \|\boldsymbol{a}\|_\nu \ .\end{aligned}$$

□

The construction of the Newton iteration is mostly identical to the analytical version.

Lemma 2.34. Let $F \in A[\boldsymbol{X}]^n$ be a system of polynomials. For all $\boldsymbol{a}, \boldsymbol{b} \in A^n$ with $\|\boldsymbol{b} - \boldsymbol{a}\|_\nu \leq \epsilon$ we have

$$\|F(\boldsymbol{b}) - F(\boldsymbol{a}) - J_F(\boldsymbol{a})(\boldsymbol{b} - \boldsymbol{a})\|_\nu \leq \epsilon^2 \ .$$

Proof. By corollary 2.32 to Taylor's formula we have

$$F(\boldsymbol{b}) = F(\boldsymbol{a}) + J_F(\boldsymbol{a}) \cdot (\boldsymbol{b} - \boldsymbol{a}) + \sum_{|\boldsymbol{\alpha}|\geq 2} \frac{1}{\boldsymbol{\alpha}!} D^{\boldsymbol{\alpha}} F(\boldsymbol{a})(\boldsymbol{b} - \boldsymbol{a})^{\boldsymbol{\alpha}}$$

and therefore

$$\|F(\boldsymbol{b}) - F(\boldsymbol{a}) - J_F(\boldsymbol{a})(\boldsymbol{b} - \boldsymbol{a})\|_\nu = \Big\| \sum_{|\boldsymbol{\alpha}|\geq 2} \frac{1}{\boldsymbol{\alpha}!} D^{\boldsymbol{\alpha}} F(\boldsymbol{a})(\boldsymbol{b} - \boldsymbol{a})^{\boldsymbol{\alpha}} \Big\|_\nu \ . \tag{2.19}$$

To prove the lemma we now look at the right side of equation (2.19). From $\|\boldsymbol{b} - \boldsymbol{a}\|_\nu \leq \epsilon$ it follows that for $i = 1, \ldots, n$, $\nu(b_i - a_i) \leq \epsilon$ and thus for all $\boldsymbol{\alpha} \in \mathbb{N}^n$ with $|\boldsymbol{\alpha}| \geq 2$

$$\nu((\boldsymbol{b} - \boldsymbol{a})^{\boldsymbol{\alpha}}) = \nu\Big(\prod_{i=1}^n (b_i - a_i)^{\alpha_i}\Big) = \prod_{i=1}^n \nu(b_i - a_i)^{\alpha_i} \leq \epsilon^2 \ .$$

Hence

$$\Big\| \sum_{|\boldsymbol{\alpha}|\geq 2} \frac{1}{\boldsymbol{\alpha}!} D^{\boldsymbol{\alpha}} F(\boldsymbol{a})(\boldsymbol{b} - \boldsymbol{a})^{\boldsymbol{\alpha}} \Big\|_\nu \leq \epsilon^2$$

and the lemma follows. □

A notable difference to the analytical case is that we can guarantee that the Jacobian of a system remains invertible.

Lemma 2.35. Let $F \in A[\boldsymbol{X}]^n$ be a system of polynomials, $\boldsymbol{a} \in A^n$ such that $\|\det(J_F(\boldsymbol{a}))\|_\nu = 1$ and $\boldsymbol{b} \in A^n$ with $\|\boldsymbol{b} - \boldsymbol{a}\|_\nu < 1$. Then

$$\|\det(J_F(\boldsymbol{b}))\|_\nu = 1 \ .$$

Proof. $\det(J_F(\boldsymbol{X}))$ is a polynomial in $A[\boldsymbol{X}]$ and therefore there exists a finite index family $I \subset \mathbb{N}^n$ such that $\det(J_F(\boldsymbol{X})) = \sum_{\boldsymbol{\alpha}\in I} c_{\boldsymbol{\alpha}} \boldsymbol{X}^{\boldsymbol{\alpha}}$ with $c_{\boldsymbol{\alpha}} \in A$. Now we have

$$\det(J_F(\boldsymbol{b})) = \sum_{\boldsymbol{\alpha}\in I} c_{\boldsymbol{\alpha}} \boldsymbol{b}^{\boldsymbol{\alpha}} = \sum_{\boldsymbol{\alpha}\in I} c_{\boldsymbol{\alpha}} \left(\boldsymbol{a} + (\boldsymbol{b}-\boldsymbol{a})\right)^{\boldsymbol{\alpha}}$$

and by the multi-index binomial theorem 2.23

$$\begin{aligned} &= \sum_{\boldsymbol{\alpha}\in I} c_{\boldsymbol{\alpha}} \Big(\sum_{\mathbf{0}\le\boldsymbol{\beta}\le\boldsymbol{\alpha}} \binom{\boldsymbol{\alpha}}{\boldsymbol{\beta}} \boldsymbol{a}^{\boldsymbol{\beta}} (\boldsymbol{b}-\boldsymbol{a})^{\boldsymbol{\alpha}-\boldsymbol{\beta}} \Big) \\ &= \underbrace{\sum_{\boldsymbol{\alpha}\in I} c_{\boldsymbol{\alpha}} \boldsymbol{a}^{\boldsymbol{\alpha}}}_{\det(J_F(\boldsymbol{a}))} + \sum_{\boldsymbol{\alpha}\in I} \sum_{\mathbf{0}\le\boldsymbol{\beta}<\boldsymbol{\alpha}} c_{\boldsymbol{\alpha}} \binom{\boldsymbol{\alpha}}{\boldsymbol{\beta}} \boldsymbol{a}^{\boldsymbol{\beta}} (\boldsymbol{b}-\boldsymbol{a})^{\boldsymbol{\alpha}-\boldsymbol{\beta}} \,. \end{aligned}$$

Since $\|\det(J_F(\boldsymbol{a}))\|_\nu = 1$ and

$$\begin{aligned} \Big\| \sum_{\boldsymbol{\alpha}\in I} \sum_{\mathbf{0}\le\boldsymbol{\beta}<\boldsymbol{\alpha}} c_{\boldsymbol{\alpha}} \binom{\boldsymbol{\alpha}}{\boldsymbol{\beta}} \boldsymbol{a}^{\boldsymbol{\beta}} (\boldsymbol{b}-\boldsymbol{a})^{\boldsymbol{\alpha}-\boldsymbol{\beta}} \Big\|_\nu &\le \max_{\boldsymbol{\alpha}\in I} \max_{\mathbf{0}\le\boldsymbol{\beta}<\boldsymbol{\alpha}} \nu \left(c_{\boldsymbol{\alpha}} \binom{\boldsymbol{\alpha}}{\boldsymbol{\beta}} \boldsymbol{a}^{\boldsymbol{\beta}} (\boldsymbol{b}-\boldsymbol{a})^{\boldsymbol{\alpha}-\boldsymbol{\beta}} \right) \\ &\le \max_{\boldsymbol{\alpha}\in I} \max_{\mathbf{0}\le\boldsymbol{\beta}<\boldsymbol{\alpha}} \nu \left((\boldsymbol{b}-\boldsymbol{a})^{\boldsymbol{\alpha}-\boldsymbol{\beta}} \right) \\ &\overset{|\boldsymbol{\alpha}-\boldsymbol{\beta}|>0 \,\wedge\, \|\boldsymbol{b}-\boldsymbol{a}\|_\nu<1}{<} 1 \end{aligned}$$

we conclude $\|\det(J_F(\boldsymbol{b}))\|_\nu = 1$ by lemma 2.7. □

Now assume that for a system of polynomials $F \in A[\boldsymbol{X}]^n$ we have an approximation $\boldsymbol{a} \in A^n$ with $\|F(\boldsymbol{a})\|_\nu \le \epsilon < 1$. If additionally the Jacobian of F, J_F, is invertible in $\boldsymbol{a}$ we get

$$\left\| (\boldsymbol{a} - J_F(\boldsymbol{a})^{-1} F(\boldsymbol{a})) - \boldsymbol{a} \right\|_\nu = \left\| -J_F(\boldsymbol{a})^{-1} F(\boldsymbol{a}) \right\|_\nu \le \|F(\boldsymbol{a})\|_\nu \le \epsilon \tag{2.20}$$

by lemma 2.33. With $\boldsymbol{b} := \boldsymbol{a} - J_F(\boldsymbol{a})^{-1} F(\boldsymbol{a})$ we have then found a suitable $\boldsymbol{b}$ for lemma 2.34 and we can formulate the following theorem:

Theorem 2.36 (Quadratic Convergence)**.** Let $F \in A[\boldsymbol{X}]^n$ be a system of polynomials, $\boldsymbol{a} \in A^n$ an approximation of F with $\|F(\boldsymbol{a})\|_\nu \le \epsilon < 1$ and $\|\det(J_F(\boldsymbol{a}))\|_\nu = 1$. Then $\boldsymbol{b} := \boldsymbol{a} - J_F(\boldsymbol{a})^{-1} F(\boldsymbol{a})$ is well defined and

$$\|\boldsymbol{b}-\boldsymbol{a}\|_\nu \le \epsilon \,,\ \|F(\boldsymbol{b})\|_\nu \le \epsilon^2 \text{ and } \|\det(J_F(\boldsymbol{b}))\|_\nu = 1 \,.$$

Proof. Since $\|\det(J_F(\boldsymbol{a}))\|_\nu = 1$ it follows by lemma 2.21 that $\det(J_F(\boldsymbol{a}))$ is a unit in A. Hence $J_F(\boldsymbol{a})$ is invertible and $\boldsymbol{b}$ well defined.

By (2.20) $\|\boldsymbol{b}-\boldsymbol{a}\|_\nu \le \epsilon < 1$ and thus $\|\det(J_F(\boldsymbol{b}))\|_\nu = 1$ by lemma 2.35. Finally we obtain by lemma 2.34

$$\begin{aligned} \epsilon^2 &\ge \|F(\boldsymbol{b}) - F(\boldsymbol{a}) - J_F(\boldsymbol{a})(\boldsymbol{b}-\boldsymbol{a})\|_\nu \\ &= \left\| F(\boldsymbol{b}) - F(\boldsymbol{a}) - J_F(\boldsymbol{a})(\boldsymbol{a} - J_F(\boldsymbol{a})^{-1}F(\boldsymbol{a}) - \boldsymbol{a}) \right\|_\nu \\ &= \left\| F(\boldsymbol{b}) - F(\boldsymbol{a}) + J_F(\boldsymbol{a}) J_F(\boldsymbol{a})^{-1} F(\boldsymbol{a}) \right\|_\nu \\ &= \|F(\boldsymbol{b})\|_\nu \,. \end{aligned} \tag{2.21}$$

□

Remark 2.37. Instead of computing $J_F(\boldsymbol{a})^{-1}$ it is sufficient to compute J^* such that $\|J^*J_F(\boldsymbol{a}) - I_n\|_\nu \leq \epsilon^2$ and to set $\boldsymbol{b} = \boldsymbol{a} - J^*F(\boldsymbol{a})$. By lemma 2.33 we have $\|(J^*J_F(\boldsymbol{a}) - I_n)F(\boldsymbol{a})\|_\nu \leq \epsilon^2$ and the inequality (2.21) would then be

$$\|F(\boldsymbol{b}) + J^*J_F(\boldsymbol{a})F(\boldsymbol{a}) - F(\boldsymbol{a})\|_\nu \leq \epsilon^2$$

and on the other hand

$$\|F(\boldsymbol{b}) + J^*J_F(\boldsymbol{a})F(\boldsymbol{a}) - F(\boldsymbol{a})\|_\nu \leq \max\{\|F(\boldsymbol{b})\|_\nu, \epsilon^2\}\,.$$

Hence $\|F(\boldsymbol{b})\|_\nu \leq \epsilon^2$.

We have seen how we can get from an approximate solution $\boldsymbol{a}$ with $\|F(\boldsymbol{a})\|_\nu \leq \epsilon < 1$ and $\|\det(J_F(\boldsymbol{a}))\|_\nu$ to a better approximation $\boldsymbol{b}$ with $\|F(\boldsymbol{b})\|_\nu \leq \epsilon^2$ and $\|\boldsymbol{b} - \boldsymbol{a}\|_\nu \leq \epsilon$. The following theorem shows that $\boldsymbol{b}$ is even unique.

Theorem 2.38 (Uniqueness)**.** Let $F \in A[\boldsymbol{X}]^n$ be a system of polynomials, $\boldsymbol{a} \in A^n$ such that $\|F(\boldsymbol{a})\|_\nu = \epsilon < 1$ and $\|\det(J_F(\boldsymbol{a}))\|_\nu = 1$. If there exists $\boldsymbol{b}$ and $\boldsymbol{b^*} \in A^n$ such that

$$\begin{array}{ll} \|\boldsymbol{b} - \boldsymbol{a}\|_\nu & \leq \epsilon \\ \|F(\boldsymbol{b})\|_\nu & \leq \epsilon^2 \end{array} \quad \text{and} \quad \begin{array}{ll} \|\boldsymbol{b^*} - \boldsymbol{a}\|_\nu & \leq \epsilon \\ \|F(\boldsymbol{b^*})\|_\nu & \leq \epsilon^2 \end{array}$$

then

$$\|\boldsymbol{b^*} - \boldsymbol{b}\|_\nu \leq \epsilon^2\,.$$

Proof. By corollary 2.32 to Taylor's formula we have

$$F(\boldsymbol{b^*}) = F(\boldsymbol{b}) + J_F(\boldsymbol{b}) \cdot (\boldsymbol{b^*} - \boldsymbol{b}) + \sum_{|\boldsymbol{\alpha}| \geq 2} \frac{1}{\boldsymbol{\alpha}!} D^{\boldsymbol{\alpha}} F(\boldsymbol{b})(\boldsymbol{b^*} - \boldsymbol{b})^{\boldsymbol{\alpha}} \tag{2.22}$$

and by lemma 2.34

$$\Big\| \sum_{|\boldsymbol{\alpha}| \geq 2} \frac{1}{\boldsymbol{\alpha}!} D^{\boldsymbol{\alpha}} F(\boldsymbol{b})(\boldsymbol{b^*} - \boldsymbol{b})^{\boldsymbol{\alpha}} \Big\|_\nu \leq \|\boldsymbol{b^*} - \boldsymbol{b}\|_\nu^2\,. \tag{2.23}$$

Since$\|\boldsymbol{b} - \boldsymbol{a}\|_\nu \leq \epsilon < 1$ it follows $\|\det(J_F(\boldsymbol{b}))\|_\nu = 1$ by lemma 2.35. Hence $\det(J_F(\boldsymbol{b}))$ is a unit in A and $J_F(\boldsymbol{b})$ is invertible. Now it follows with (2.22)

$$\begin{aligned} \|\boldsymbol{b^*} - \boldsymbol{b}\|_\nu &= \Big\| J_F(\boldsymbol{b})^{-1} \Big(F(\boldsymbol{b^*}) - F(\boldsymbol{b}) - \sum_{|\boldsymbol{\alpha}| \geq 2} \frac{1}{\boldsymbol{\alpha}!} D^{\boldsymbol{\alpha}} F(\boldsymbol{b})(\boldsymbol{b^*} - \boldsymbol{b})^{\boldsymbol{\alpha}} \Big) \Big\|_\nu \\ &\leq \Big\| F(\boldsymbol{b^*}) - F(\boldsymbol{b}) - \sum_{|\boldsymbol{\alpha}| \geq 2} \frac{1}{\boldsymbol{\alpha}!} D^{\boldsymbol{\alpha}} F(\boldsymbol{b})(\boldsymbol{b^*} - \boldsymbol{b})^{\boldsymbol{\alpha}} \Big\|_\nu \\ &\leq \max\Big\{ \|F(\boldsymbol{b^*})\|_\nu, \|F(\boldsymbol{b})\|_\nu, \Big\| \sum_{|\boldsymbol{\alpha}| \geq 2} \frac{1}{\boldsymbol{\alpha}!} D^{\boldsymbol{\alpha}} F(\boldsymbol{b})(\boldsymbol{b^*} - \boldsymbol{b})^{\boldsymbol{\alpha}} \Big\|_\nu \Big\} \\ &\overset{(2.23)}{\leq} \max\Big\{ \epsilon^2, \epsilon^2, \|\boldsymbol{b^*} - \boldsymbol{b}\|_\nu^2 \Big\}\,. \end{aligned}$$

Since $\|\boldsymbol{b^*} - \boldsymbol{b}\|_\nu = \|\boldsymbol{b^*} - \boldsymbol{a} - (\boldsymbol{b} - \boldsymbol{a})\|_\nu \leq \max\{\|\boldsymbol{b^*} - \boldsymbol{a}\|_\nu, \|\boldsymbol{b} - \boldsymbol{a}\|_\nu\} \leq \epsilon < 1$ it follows that $\|\boldsymbol{b^*} - \boldsymbol{b}\|_\nu^2 < \|\boldsymbol{b^*} - \boldsymbol{b}\|_\nu$. Therefore we conclude

$$\|\boldsymbol{b^*} - \boldsymbol{b}\|_\nu \leq \max\left\{\epsilon^2, \epsilon^2, \|\boldsymbol{b^*} - \boldsymbol{b}\|_\nu^2\right\} \leq \epsilon^2$$

□

The previous results combined now yield the Newton iteration.

Theorem 2.39 (Newton iteration)**.** Let $F \in A[\boldsymbol{X}]^n$ be a system of polynomials, $\boldsymbol{a}^{(0)} \in A^n$ such that $\left\|F(\boldsymbol{a}^{(0)})\right\|_\nu \leq \epsilon < 1$ and $\left\|\det(J_f(\boldsymbol{a}^{(0)}))\right\|_\nu = 1$. Define the sequence $(\boldsymbol{a}^{(k)})$ by

$$\boldsymbol{a}^{(j+1)} := \boldsymbol{a}^{(j)} - J_F(\boldsymbol{a}^{(j)})^{-1} F(\boldsymbol{a}^{(j)}), \quad j \geq 0\,.$$

For all positive integers k we have

$$\left\|F(\boldsymbol{a}^{(k)})\right\|_\nu \leq \epsilon^{2^k}, \quad \left\|\det(J_F(\boldsymbol{a}^{(k)}))\right\|_\nu = 1 \quad \text{and} \quad \left\|\boldsymbol{a}^{(k)} - \boldsymbol{a}^{(0)}\right\|_\nu \leq \epsilon\,. \tag{2.24}$$

Furthermore, $\boldsymbol{a}^{(k)}$ is unique, i.e., for all $\boldsymbol{b} \in A^n$ with $\|F(\boldsymbol{b})\|_\nu \leq \epsilon^{2^k}$ and $\left\|\boldsymbol{b} - \boldsymbol{a}^{(0)}\right\|_\nu \leq \epsilon$ we have

$$\left\|\boldsymbol{b} - \boldsymbol{a}^{(k)}\right\|_\nu \leq \epsilon^{2^k}\,. \tag{2.25}$$

Proof. The statement (2.24) follows immediately by induction over k by theorem 2.36 and (2.25) by theorem 2.38. □

Remark 2.40. This algebraic version of the Newton iteration is even stronger than the analytical version since we can guarantee that the Jacobian remains invertible!

Since the proof was constructive we can formulate the following Newton iteration algorithm:

Algorithm 2.41 Newton iteration

Input: $F \in A[X_1, \ldots, X_n]^n$ and its Jacobian $J_F \in A[X_1, \ldots, X_n]^{n,n}$, $\boldsymbol{a}^{(0)} \in A^n$ such that $\left\|F(\boldsymbol{a}^{(0)})\right\|_\nu \leq \epsilon < 1$ and $\left\|\det(J_F(\boldsymbol{a}^{(0)}))\right\|_\nu = 1$ and $D \in \mathbb{N}$.
Output: $\boldsymbol{a} \in A^n$ such that $\|F(\boldsymbol{a})\|_\nu \leq \epsilon^D$ and $\|\det(J_F(\boldsymbol{a}))\|_\nu = 1$

```
r := ⌈log_2 D⌉
for k := 1, ..., r do
    Compute J^(k) ∈ A^{n,n} such that ‖J^(k) J_F(a^(k-1)) − I_n‖_ν ≤ ε^{2^k}
    a^(k) := a^(k-1) − J^(k) F(a^(k-1))
end for
return a^(r)
```

Theorem 2.42. The Newton iteration algorithm 2.41 works correctly, its output is unique and it needs at most $\mathcal{O}((\log_2(D) + 1)n^3)$ arithmetic operations.

Proof. The correctness and uniqueness follows by theorem 2.39 and remark 2.37. The dominant step for the complexity is the computation of $J^{(k)}$. The cost to compute $J^{(k)}$ is bounded by the number of operations necessary to compute the inversion of $J_F(\boldsymbol{a}^{(k)})$. Since this needs at most $\mathcal{O}(n^3)$ arithmetic operations the statement follows. □

Finally we state that if $\hat{A}$ is a complete Noetherian local ring and ν the $\mathfrak{m}$-adic valuation, the Newton iteration always converges to a unique limit.

Corollary 2.43. Let $\hat{A}$ be a complete Noetherian local ring with maximal ideal $\mathfrak{m}$ and equipped with the $\mathfrak{m}$-adic valuation. If for a system of polynomials $F \in \hat{A}[\boldsymbol{X}]^n$ an $\boldsymbol{a} = (a_i) \in \hat{A}^n$ exists such that

$$\|F(\boldsymbol{a})\|_{\mathfrak{m}} < 1 \quad \text{and} \quad \|\det(J_F(\boldsymbol{a}))\|_{\mathfrak{m}} = 1$$

then there exists an unique $\hat{\boldsymbol{a}} \in \hat{A}^n$ such that

$$F(\hat{\boldsymbol{a}}) = 0\,, \quad \det(J_F(\hat{\boldsymbol{a}})) \text{ unit} \quad \text{and} \quad \|\hat{\boldsymbol{a}} - \boldsymbol{a}\|_{\mathfrak{m}} < 1\,.$$

Proof. With $\boldsymbol{a}^{(0)} := \boldsymbol{a}$ the sequence

$$\boldsymbol{a}^{(k+1)} := \boldsymbol{a}^{(k)} - J_F(\boldsymbol{a}^{(k)})^{-1} F(\boldsymbol{a}^{(k)}) \quad , \quad k \geq 0 \tag{2.26}$$

is well defined and a Cauchy sequence as, by theorem 2.39, for $i = 1, \ldots, n$ and all integers N

$$a_i^{(j)} - a_i^{(k)} \in \mathfrak{m}^N \text{ for all } j, k > \lceil \log_2 N \rceil.$$

Define $\hat{\boldsymbol{a}} := \lim \boldsymbol{a}^{(i)} \in \hat{A}$. By theorem 2.39 $\hat{\boldsymbol{a}}$ is unique, $F(\hat{\boldsymbol{a}}) = 0$, $\det(J_F(\hat{\boldsymbol{a}}))$ a unit and $\|\hat{\boldsymbol{a}} - \boldsymbol{a}\|_{\mathfrak{m}} < 1$. □

Remark 2.44. Remember that $\|F(\boldsymbol{a})\|_{\mathfrak{m}} < 1$ if and only if $f_i(\boldsymbol{a}) \equiv 0 \bmod \mathfrak{m}$ for $1 \leq i \leq n$ and $\|\det(J_F(\boldsymbol{a}))\|_{\mathfrak{m}} = 1$ if and only if $\det(J_F(\boldsymbol{a}))$ is a unit.

2.4 Sylvester matrix and resultant

This section is based on Chapter 6 in Modern Computer Algebra [19]. Let us now take a look at our original problem. We want to factor the polynomial

$$f(X, Y) = Y^3 + (X - 1)Y^2 + (-X + 1)Y - 1 \in \mathbb{Q}[X][Y]$$

and already derived that this is equivalent to a solution $g_1, g_0, h_0 \in \mathbb{Q}[X]$ such that

$$F(h_0, g_1, g_0) = \begin{bmatrix} g_1 + h_0 - (X - 1) \\ g_1 h_0 + g_0 - (-X + 1) \\ g_0 h_0 - (-1) \end{bmatrix} = \mathbf{0}\,.$$

Since $f(0, Y) = (Y^2 + 1)(Y - 1)$ we have

$$F(-1, 0, 1) = \begin{bmatrix} -X \\ X \\ 0 \end{bmatrix}\,.$$

If we now interpret f, g_1, g_0 and h_0 as elements of $\mathbb{Q}[[X]]$ F is a system over the complete Noetherian local ring $\mathbb{Q}[[X]]$ with maximal ideal $\mathfrak{m} = (X)$. Then $\|F(-1, 0, 1)\|_{\mathfrak{m}} = 1/2$ and if $J_F(-1, 0, 1)$ is invertible we can obtain a solution for F by Newton iteration.

Consider the Jacobian of F

$$J_F(h_0, g_1, g_0) = \begin{bmatrix} 1 & 1 & 0 \\ g_1 & h_0 & 1 \\ g_0 & 0 & h_0 \end{bmatrix} \quad \text{and in particular} \quad J_F(-1, 0, 1) = \begin{bmatrix} 1 & 1 & 0 \\ 0 & -1 & 1 \\ 1 & 0 & -1 \end{bmatrix} .$$

$J_F(-1, 0, 1)$ is invertible and we can apply the Newton iteration to obtain a unique solution. Since our initial approximation is a factorization of $f(0, Y)$ can we maybe give a general condition such that the Jacobian of our initial solution is invertible?

Let k be a field and $g, h \in k[X]$ univariate polynomials with $\deg(g) = n$ and $\deg(h) = m$. Then for $d \in \mathbb{N}$

$$k_{<d}[X] := \{f \in k[X] \mid \deg(f) < d\}$$

is the vector space of polynomials over k with degree less than d. We define the linear-combination map as

$$\varphi_{g,h} : k_{<m}[X] \times k_{<n}[X] \to k_{<n+m}[X], \quad (s, t) \mapsto sg + th \, . \tag{2.27}$$

Since $\varphi_{g,h}$ is a linear mapping between vector spaces there exists a transformation matrix of $\varphi_{g,h}$, which we now want to determine. Choose $\{X^{m-1}, \ldots, X, 1\}$ as a monomial basis for $k_{<m}[X]$ and analog bases for $k_{<n}[X]$ and $k_{<n+m}[X]$. Consider at first the mapping

$$\varphi_g : k_{<m}[X] \to k_{<n+m}[X], \; s \mapsto sg \, .$$

Let s be in $k_{<m}[X]$ with $s = \sum_{i=0}^{m-1} s_i X^i$ and write $g = \sum_{j=0}^{n} g_j X^j$. Then

$$sg = \left(\sum_{i=0}^{m-1} s_i X^i \right) g = \sum_{i=0}^{m-1} s_i \left(g X^i \right)$$

and if we interpret g as an element in $k_{<n+m}[X]$ it has the coordinate vector

$$[0, \ldots, 0, g_n, \ldots, g_0]^T \in k^{n+m}$$

and gX^i, $1 \leq i < m$,

$$[\underbrace{0, \ldots, 0}_{m-1-i}, g_n, \ldots, g_0, \underbrace{0, \ldots, 0}_{i}]^T \in k^{n+m}$$

Now we can obtain the $(n+m)\times m$ transformation matrix of φ_g by

$$\begin{bmatrix} g_n & 0 & \dots & 0 \\ g_{n-1} & g_n & \ddots & \vdots \\ \vdots & g_{n-1} & \ddots & 0 \\ \vdots & \ddots & \ddots & g_n \\ \vdots & \ddots & \ddots & g_{n-1} \\ g_1 & \ddots & \ddots & \vdots \\ g_0 & g_1 & \ddots & \vdots \\ 0 & g_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & g_1 \\ 0 & \dots & 0 & g_0 \end{bmatrix} .$$

We can analogous obtain the $(n+m)\times n$ transformation matrix of

$$\varphi_h : k_{<n}[X] \to k_{<n+m}[X],\ t \mapsto th\ .$$

The transformation matrix of $\varphi_{g,h}$ is thus

$$\begin{bmatrix} g_n & 0 & \dots & 0 & h_m & 0 & \dots & \dots & 0 \\ g_{n-1} & g_n & \ddots & \vdots & h_{m-1} & h_m & \ddots & & \vdots \\ \vdots & g_{n-1} & \ddots & 0 & \vdots & h_{m-1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & g_n & \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & g_{n-1} & \vdots & \ddots & \ddots & \ddots & h_m \\ g_2 & \ddots & \ddots & \vdots & h_1 & \ddots & \ddots & \ddots & h_{m-1} \\ g_1 & g_2 & \ddots & \vdots & h_0 & h_1 & \ddots & \ddots & \vdots \\ g_0 & g_1 & \ddots & \vdots & 0 & h_0 & \ddots & \ddots & \vdots \\ 0 & g_0 & \ddots & g_2 & \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & g_1 & \vdots & & \ddots & \ddots & h_1 \\ 0 & \dots & 0 & g_0 & 0 & \dots & \dots & 0 & h_0 \end{bmatrix} . \tag{2.28}$$

For the construction of (2.28) we did not use the fact that k is a field. Therefore we can give the following more general definition.

Definition 2.45 (Sylvester matrix and resultant)**.** Let A be a commutative ring and $g, h \in A[X]$ two polynomials with $g = \sum_{i=0}^{n} g_i X^i$ and $h = \sum_{i=0}^{m} h_i X^i$. Then the $(n+m)\times(n+m)$ matrix (2.28) is the *Sylvester matrix* of g and h denoted by $\mathrm{Syl}(g,h)$. The determinant of $\mathrm{Syl}(g,h)$ is called the *resultant* of g and h denoted by $\mathrm{res}(g,h)$.

Example 2.46. We continue our example. Consider the initial factorization

$$f(0,Y) = Y^3 - Y^2 + Y - 1 = (Y^2+1)(Y-1) \in \mathbb{Q}[[X]][Y] .$$

Then

$$\mathrm{Syl}(Y^2+1, Y-1) = \begin{bmatrix} 1 & 1 & 0 \\ 0 & -1 & 1 \\ 1 & 0 & -1 \end{bmatrix} \quad \text{and} \quad \mathrm{res}(Y^2+1, Y-1) = 2\ .$$

Notice that $\mathrm{Syl}(Y^2+1, Y-1)$ and $J_F(-1, 0, 1)$ coincide!

Let g and h be univariate polynomials over a (commutative) ring A. With the Sylvester matrix we can reformulate the linear-combination mapping $\varphi_{g,h}$ as the corresponding linear transformation mapping:

$$\Phi_{g,h} : A^m \times A^n \to A^{n+m}, \quad \left(\begin{bmatrix} s_{m-1} \\ \vdots \\ s_0 \end{bmatrix}, \begin{bmatrix} t_{n-1} \\ \vdots \\ t_0 \end{bmatrix} \right) \mapsto \mathrm{Syl}(g,h) \begin{bmatrix} s_{m-1} \\ \vdots \\ s_0 \\ t_{n-1} \\ \vdots \\ t_0 \end{bmatrix} \tag{2.29}$$

With the linear combination mapping $\varphi_{g,h}$ and the Sylvester matrix we can now prove (based on [19]) the following astonishing theorem which links the questions whether g and h are strongly relatively prime to the resultant of g and h.

Theorem 2.47. Let A be a (commutative) ring and $g, h \in A[X]$ univariate polynomials with $g = \sum_{i=0}^{n} g_i X^i$ and $h = \sum_{i=0}^{m} h_i X^i$ such that g_n and h_m are units in A. Let $\varphi_{g,h}$ be the linear combination mapping

$$\varphi_{g,h} : A_{<m}[X] \times A_{<n}[X] \to A_{<n+m}[X], \quad (s,t) \mapsto sg + th.$$

Then the following statements are equivalent:

(1) There exists $(s,t) \in A_{<m}[X] \times A_{<n}[X]$ such that $sg + th = 1$

(2) $\varphi_{g,h}$ is an isomorphism

(3) $\mathrm{res}(g,h)$ is a unit in A

Proof. Let $\Phi_{g,h}$ be defined as in (2.29). Then

$$\begin{aligned} \varphi_{g,h} \text{ isomorphism} &\iff \Phi_{g,h} \text{ isomorphism} \\ &\iff \mathrm{Syl}(g,h) \text{ invertible} \\ &\iff \mathrm{res}(g,h) = \det(\mathrm{Syl}(g,h)) \text{ unit in A}\ . \end{aligned}$$

This shows the equivalence of (2) and (3). Now let $\varphi_{g,h}$ be an isomorphism. Then $\varphi_{g,h}^{-1}(1) = (s,t)$ such that $sg + th = 1$. Hence (2) implies (1).

Finally let $(s,t) \in A_{<m}[X] \times A_{<n}[X]$ such that $sg + th = 1$. We claim that then exists $(s_k, t_k) \in A_{<m}[X] \times A_{<n}[X]$ such that $s_k g + t_k h = X^k$ for $0 \le k < n + m$. We prove this by induction over k. For $k = 0$ set $s_0 = s$ and $t_0 = t$. Now assume the induction hypothesis holds for some $k - 1 < n + m - 1$. Then there exists $(s_{k-1}, t_{k-1}) \in A_{<m}[X] \times A_{<n}[X]$ such that $s_{k-1} g + t_{k-1} h = X^{k-1}$ and

$$X^k = (s_{k-1} g + t_{k-1} h) X = s_{k-1} X g + t_{k-1} X h\ .$$

Since h_m is a unit there exists $a \in A$ such that $s_{k-1}X = ah + s_k$ with $\deg(s_k) < m$. Then

$$\begin{aligned} X^k &= s_{k-1}Xg + t_{k-1}Xh \\ &= s_{k-1}Xg + t_{k-1}Xh - ahg + ahg \\ &= (s_{k-1}X - ah)g + (t_{k-1}X + ag)h \\ &= s_k g + (t_{k-1}X + ag)h \end{aligned}$$

Now $\deg(s_k g) < n+m$ and $\deg((t_{k-1}X+ag)h) \leq \deg(t_{k-1}X+ag)+m$. Since $k < n+m$ is $\deg(t_{k-1}X+ag)+m < n+m$ and hence $\deg(t_{k-1}X+ag) < n$. With $t_k := t_{k-1}X+ag$ it follows the hypothesis.

As a result there exists for $0 \leq k < n+m$ $(s_k, t_k) \in A_{<m}[X] \times A_{<n}[X]$ such that $\varphi_{g,h}(s_k, t_k) = X^k$ and therefore is $\varphi_{g,h}$ surjective and hence an isomorphism which completes the proof. □

Remark 2.48. The equivalence of (2) and (3) is valid for any $g, h \in A[X]$.

Definition 2.49. Let A be a ring and g and $h \in A[X]$ polynomials with $\deg(g) = n$ and $\deg(h) = m$. If there exists $s \in A_{<m}[X]$ and $t \in A_{<n}[X]$ such that $sg + th = 1$ we call g and h *strongly relatively prime*.

Remark 2.50. If $A[X]$ is a unique factorization domain and g and h are strongly relatively prime then g and h are also *relatively prime* in the sense that $\gcd(g, h) = 1$. Suppose this were not the case, then $\gcd(g, h)$ divides $1 = sg + th$. Since $\gcd(g, h)$ it not a unit and A an integral domain this is a contradiction. If $A[X]$ is a principal ideal domain then, by Bezout's identity, relatively prime g and h are also strongly relatively prime.

We can also obtain the following useful corollary

Corollary 2.51. Let A be an unique factorization domain and $g, h \in A[X]$. Then $\operatorname{res}(g, h) \neq 0$ if and only if $\gcd(g, h)$ constant.

Proof. Let k be the field of fractions of A. Then $k[X]$ is a principal ideal domain. Hence g, h are strongly relatively prime in k if and only if $\gcd(g, h) = 1$ in k by Bezout's identity. Thus $\operatorname{res}(g, h)$ is a unit, i.e. $\operatorname{res}(g, h) \neq 0$, if and only if $\gcd(g, h) = 1$ in k by theorem 2.47. Since $\gcd(g, h) = 1$ in k if and only if $\gcd(g, h)$ constant in A we conclude the proof. □

With the resultant we have a powerful (theoretical) tool to determine whether two polynomials are (strongly) relatively prime. In the following we are particularly interested whether polynomials $g, h \in A[X]$ remain (strongly) relatively prime in $A/\mathfrak{m}$ for some proper ideal $\mathfrak{m}$. Since $\operatorname{res}(g, h)$ is a polynomial in $A[X]$ one might assume that there is no difference whether we first include g, h in $A/\mathfrak{m}$ and then compute the resultant or include $\operatorname{res}(g, h)$ in $A/\mathfrak{m}$. But consider the following example

Example 2.52. For $g = Y^2X^3 - X$ and $h = YX + 1 \in \mathbb{Q}[Y][X]$ and $\mathfrak{m} = (Y)$ we have

$$\operatorname{Syl}(g, h) = \begin{bmatrix} Y^2 & Y & 0 & 0 \\ 0 & 1 & Y & 0 \\ 0 & 0 & 1 & Y \\ -1 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \operatorname{Syl}(\bar{g}, \bar{h}) = \operatorname{Syl}(-X, 1) = \begin{bmatrix} -1 \end{bmatrix} .$$

Hence $\operatorname{res}(g,h) = Y^2(Y+1) \equiv 0 \bmod Y$ but $\operatorname{res}(\bar{g},\bar{h}) \equiv -1 \bmod Y$.

The reason is that the Sylvester matrices are rather different. Thus we have to find a sufficient condition that ensures that the resultants coincide. But at first we need some notation.

Definition 2.53. For a polynomial $f \in A[X]$ denote by $\operatorname{lt}(f)$ the *leading term* of f, i.e. the term with highest degree of f. Moreover, denote by $\operatorname{lc}(f)$ the *leading coefficient* of f, i.e. the coefficient of the monomial of $\operatorname{lt}(f)$.

Lemma 2.54. Let A be a integral domain, $\mathfrak{m} \subset A$ a proper ideal and $g, h \in A[X]$ non-zero polynomials. If $\operatorname{lc}(g)$ and $\operatorname{lc}(h)$ are units in A then

$$\operatorname{res}(g \bmod \mathfrak{m}, h \bmod \mathfrak{m}) \equiv \operatorname{res}(g,h) \bmod \mathfrak{m} .$$

Proof. Since $\operatorname{lc}(g)$ and $\operatorname{lc}(h)$ are units in A it follows that $\deg(g) = \deg(g \bmod \mathfrak{m})$ and $\deg(h) = \deg(h \bmod \mathfrak{m})$. The construction of $\operatorname{Syl}(g,h)$ depends on the degree of g and h. Thus $\operatorname{Syl}(g,h)$ and $\operatorname{Syl}(g \bmod \mathfrak{m}, h \bmod \mathfrak{m})$ have the same size. Since the resultant is a polynomial in the coefficients of g and h the statement follows. □

For the case that A is an Noetherian domain we can now give the following useful connection between (strongly) relatively prime polynomials.

Corollary 2.55. Let A be a Noetherian domain with maximal ideal $\mathfrak{m}$, $\hat{A}_{\mathfrak{m}}$ its completion and $g, h \in A[X]$ non-zero polynomials with $\operatorname{lc}(g)$ and $\operatorname{lc}(h)$ not in $\mathfrak{m}$. Then the following statements are equivalent:

(1) $\gcd(g \bmod \mathfrak{m}, h \bmod \mathfrak{m}) \notin \mathfrak{m}$

(2) g and h are relatively prime in $A/\mathfrak{m}[X]$

(3) g and h are strongly relatively prime in $A/\mathfrak{m}^k[X]$ for all $k \in \mathbb{N}_{\geq 1}$

(4) g and h are strongly relatively prime in $\hat{A}_{\mathfrak{m}}[X]$

Proof. Since $\mathfrak{m}$ is a maximal ideal, $A/\mathfrak{m}$ is a field and thus $\operatorname{lc}(g)$ and $\operatorname{lc}(h)$ are units in $A/\mathfrak{m}$. This also implies the equivalence of (1) and (2).

Moreover, $A/\mathfrak{m}[X]$ is a principal ideal domain. Therefore g and h are relatively prime in $A/\mathfrak{m}$ if and only if g and h are strongly relatively prime in $A/\mathfrak{m}$ by Bezout's identity. By theorem 2.47 it follows that g and h are relatively prime in $A/\mathfrak{m}$ if and only if $\operatorname{res}(g,h)$ is a unit in $A/\mathfrak{m}$. Let k be a positive integer. By lemma 2.54 we have $\operatorname{res}(g \bmod \mathfrak{m}^k, h \bmod \mathfrak{m}^k) \equiv \operatorname{res}(g,h) \bmod \mathfrak{m}^k$ and by lemma 2.12 $\operatorname{res}(g,h)$ unit in $A/\mathfrak{m}$ if and only if $\operatorname{res}(g,h)$ unit in $A/\mathfrak{m}^k$. This combined implies the equivalence of (2) and (3). From the construction of $\hat{A}_{\mathfrak{m}}$ it follows immediately the equivalence of (3) and (4). □

2.5 Hensel lifting

The content of this section was derived on my own. In example 2.46 we have seen that the Jacobian of our initial approximation and the Sylvester matrix of the polynomials of the corresponding factorization of $f(0,Y)$ coincide. This is in fact always the case and therefore the Jacobian is invertible if and only if the polynomials of the corresponding factorization are strongly relatively prime by our previous result. This can be seen as follows:

Let $\hat{A}$ be a complete Noetherian local ring with maximal ideal $\mathfrak{m}$, $f \in \hat{A}[X]$ a monic polynomial and write $f = \sum_{i=0}^{d-1} f_i X^i + X^d$. Let m and n be positives integers such that $n+m=d$. Assume we want to determine polynomials $g = \sum_{i=0}^{n-1} g_i X^i + X^n \in \hat{A}[X]$ and $h = \sum_{i=0}^{m-1} h_i X^i + X^m \in \hat{A}[X]$ such that

$$f = \sum_{i=0}^{d-1} f_i X^i + X^d = \Big(\sum_{i=0}^{n-1} g_i X^i + X^n\Big)\Big(\sum_{i=0}^{m-1} h_i X^i + X^m\Big) = gh\ .$$

This yields $n+m$ equations (with $g_n = 1$ and $h_m = 1$ for convenience)

$$f_k = \sum_{i+j=k} g_i h_j\ , \quad 0 \le k < n+m$$

which is equivalent to finding roots for $n+m$ polynomials

$$F_k := \sum_{i+j=k} G_i H_j - f_k \in \hat{A}[\boldsymbol{H}, \boldsymbol{G}]\ , \quad 0 \le k < n+m$$

where $(\boldsymbol{H}, \boldsymbol{G}) := (H_{m-1}, \ldots, H_0, G_{n-1}, \ldots, G_0)$ is a family of indeterminates and $G_n = 1$ and $H_n = 1$. Now we can define

$$F := \begin{bmatrix} F_{m+n-1} \\ \vdots \\ F_0 \end{bmatrix} \in \hat{A}[\boldsymbol{H}, \boldsymbol{G}]^{m+n}\ . \tag{2.30}$$

Thus the problem of determining monic polynomials $g \in \hat{A}[X]$, $\deg(g) = n$, and $h \in \hat{A}[X]$, $\deg(h) = m$, such that $f = gh$ is equivalent to determining $(\boldsymbol{h}, \boldsymbol{g}) \in \hat{A}^{m+n}$ such that

$$F(\boldsymbol{h}, \boldsymbol{g}) = \boldsymbol{0}\ .$$

Now assume we have an approximate solution $(\boldsymbol{h}, \boldsymbol{g}) \in \hat{A}^{m+n}$ with $\|F(\boldsymbol{h}, \boldsymbol{g})\|_{\mathfrak{m}} < 1$. By corollary 2.43 we can obtain a unique solution to (2.30) via Newton iteration if $J_F(\boldsymbol{h}, \boldsymbol{g})$ is invertible. Therefore we take a closer look at the Jacobian

$$J_F(\boldsymbol{H}, \boldsymbol{G}) = \begin{bmatrix} \left(\frac{\partial F_{m+n-1}}{\partial H_i}\right)_{m>i\ge 0} & \left(\frac{\partial F_{m+n-1}}{\partial G_j}\right)_{n>j\ge 0} \\ \vdots & \vdots \\ \left(\frac{\partial F_0}{\partial H_i}\right)_{m>i\ge 0} & \left(\frac{\partial F_0}{\partial G_j}\right)_{n>j\ge 0} \end{bmatrix}\ .$$

For $0 \leq k < m+n$ and $0 \leq i < m$ we have

$$\begin{aligned}\frac{\partial F_k}{\partial H_i} &= \frac{\partial}{\partial H_i}\left(\sum_{i+j=k} G_i H_j - f_k\right) \\ &= \sum_{i+j=k} \frac{\partial(G_i H_j)}{\partial H_i} \\ &= \sum_{j=\max\{0,-n+k\}}^{\min\{m,k\}} \frac{\partial(G_{k-j}H_j)}{\partial H_i} = \begin{cases} G_{k-i} & , \ \max\{0,-n+k\} \leq i \leq \min\{m,k\} \\ 0 & , \ \text{else} \end{cases}\end{aligned}$$

and by an analogous computation for $0 \leq j < n$

$$\frac{\partial F_k}{\partial G_j} = \begin{cases} H_{k-j} & , \ \max\{0,-m+k\} \leq j \leq \min\{n,k\} \\ 0 & , \ \text{else} \end{cases}.$$

In particular

$$\begin{bmatrix} \frac{\partial F_{m+n-1}}{\partial H_i} \\ \vdots \\ \frac{\partial F_0}{\partial H_i} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ G_{n-1} \\ \vdots \\ G_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \\ \\ \\ \Big\} m-1-i \\ \\ \\ \\ \\ \Big\} i \end{matrix} \quad \text{and} \quad \begin{bmatrix} \frac{\partial F_{m+n-1}}{\partial G_j} \\ \vdots \\ \frac{\partial F_0}{\partial G_j} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ H_{m-1} \\ \vdots \\ H_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \\ \\ \\ \Big\} n-1-j \\ \\ \\ \\ \\ \Big\} j \end{matrix}$$

Hence

$$J_F(\boldsymbol{h}, \boldsymbol{g}) = \mathrm{Syl}(g,h) \tag{2.31}$$

and thus J_F is invertible if and only if $\mathrm{res}(g,h)$ is a unit in $\hat{A}$. To obtain a unique solution for F it is therefore sufficient that we have strongly relatively prime polynomials $g, h \in \hat{A}[X]$ such that $f \equiv gh \bmod \mathfrak{m}$ by theorem 2.47!

Proposition 2.56 (Hensel lifting)**.** Let $\hat{A}$ be a complete Noetherian local ring with maximal ideal $\mathfrak{m}$ and $f \in \hat{A}[X]$ a non-zero polynomial with $\mathrm{lc}(f) \notin \mathfrak{m}$. If there exists $g, h \in \hat{A}[X]$ such that g and h are strongly relatively prime and

$$f \equiv gh \bmod \mathfrak{m}$$

then there exists unique strongly relatively prime polynomials $\hat{g}$ and $\hat{h} \in \hat{A}[X]$ such that

$$f = \hat{g}\hat{h} \quad \text{and} \quad \begin{matrix} \hat{g} \equiv g \bmod \mathfrak{m} \\ \hat{h} \equiv h \bmod \mathfrak{m} \end{matrix}. \tag{2.32}$$

Proof. Since $\mathrm{lc}(f) \notin \mathfrak{m}$ we have also $\mathrm{lc}(g) \notin \mathfrak{m}$ and $\mathrm{lc}(h) \notin \mathfrak{m}$. Thus $\mathrm{lc}(f), \mathrm{lc}(g)$ and $\mathrm{lc}(h)$ are units in $\hat{A}$. Therefore we assume without loss of generality that f, g and h are monic (see remark 2.58 for details) and write $g = \sum_{i=0}^{n-1} g_i X^i + X^n$ and $h = \sum_{i=0}^{m-1} h_i X^i + X^m$.

Now we can construct system (2.30) with $F \in \hat{A}[\boldsymbol{H}, \boldsymbol{G}]^{m+n}$. As $f \equiv gh \bmod \mathfrak{m}$, F has an approximate solution $(\boldsymbol{h}, \boldsymbol{g}) := (h_{m-1}, \ldots, h_0, g_{n-1}, \ldots, g_0) \in \hat{A}^{m+n}$ with $\|F(\boldsymbol{h}, \boldsymbol{g})\|_{\mathfrak{m}} < 1$ and $J_F(\boldsymbol{h}, \boldsymbol{g}) = \mathrm{Syl}(g, h)$ by (2.31). Furthermore, g and h are strongly relatively prime and thus $\mathrm{res}(g, h) = \det(J_F(\boldsymbol{h}, \boldsymbol{g}))$ is a unit by theorem 2.47. Thus all conditions for the corollary 2.43 to the Newton iteration are satisfied. Therefore there exists a unique $(\hat{\boldsymbol{h}}, \hat{\boldsymbol{g}}) = (\hat{h}_{m-1}, \ldots, \hat{h}_0, \hat{g}_{n-1}, \ldots, \hat{g}_0) \in \hat{A}^{m+n}$ such that

$$F(\hat{\boldsymbol{h}}, \hat{\boldsymbol{g}}) = \boldsymbol{0},\ \det(J_f(\hat{\boldsymbol{h}}, \hat{\boldsymbol{g}})) \text{ unit and } \|(\hat{\boldsymbol{h}}, \hat{\boldsymbol{g}}) - (\boldsymbol{h}, \boldsymbol{g})\|_{\mathfrak{m}} < 1.$$

Define the polynomials

$$\hat{g} := \sum_{i=0}^{n-1} \hat{g}_i X^i + X^n \qquad \text{and} \qquad \hat{h} := \sum_{i=0}^{m-1} \hat{h}_i X^i + X^m \ .$$

Then $\hat{g}$ and $\hat{h}$ satisfy (2.32) and since $\det(J_f(\hat{\boldsymbol{h}}, \hat{\boldsymbol{g}})) = \mathrm{res}(\hat{g}, \hat{h})$, $\hat{g}$ and $\hat{h}$ are strongly relatively prime by theorem 2.47. □

Remark 2.57. We call the polynomials $\hat{g}$ and $\hat{h}$ *lifted*.

Remark 2.58. Since $\mathrm{lc}(f)$, $\mathrm{lc}(g)$ and $\mathrm{lc}(h)$ are units we can write $\mathrm{lc}(f)f' = f$, $\mathrm{lc}(g)g' \equiv g \bmod \mathfrak{m}$ and $\mathrm{lc}(h)h' \equiv h \bmod \mathfrak{m}$ with monic polynomials f', g' and h'. It is therefore sufficient to apply our lifting procedure to the monic polynomials such that we obtain lifted polynomial $\hat{g}'$ and $\hat{h}'$ and then to return $\mathrm{lc}(g)\hat{g}$ and $\mathrm{lc}(h)\hat{h}$.

For the rest of this section let $f \in \hat{A}[X]$ be a monic polynomial with $\deg(f) = n$. We have seen that it is possible to lift a factorization $f \equiv gh \bmod \mathfrak{m}$ with strongly relatively prime polynomials $g, h \in \hat{A}[X]$ *uniquely* to a factorization $f = \hat{g}\hat{h}$ with $\hat{g}, \hat{h} \in \hat{A}[X]$. To be useful in practice we have to generalize the statement in such a way that we can lift strongly relatively prime monic polynomials $g_1, \ldots, g_r \in \hat{A}[X]$ with $\deg(g_i) = n_i$ and

$$f \equiv \prod_{i=1}^{r} g_i^{e_i} \bmod \mathfrak{m},\ e_i \geq 1 \tag{2.33}$$

to *unique* polynomials $\hat{g}_1, \ldots, \hat{g}_r \in \hat{A}[X]$ such that

$$f = \textstyle\prod_{i=1}^{r} \hat{g}_i^{e_i} \quad \text{and} \quad \hat{g}_i \equiv g_i \bmod \mathfrak{m}\ ,\ i = 1, \ldots, r\ .$$

We make the generalization in two steps. First we consider the case that $e_i = 1$ for $i = 1, \ldots, n$ and then the general case (2.33).

Therefore suppose that $f \equiv \prod_{i=1}^{r} g_i \bmod \mathfrak{m}$ with $g_i = \sum_{j=0}^{n_i-1} g_{i,j} X^j + X^{n_i}$. If we define the families of indeterminates $\boldsymbol{G}_i := (G_{i,n_i-1}, \ldots, G_{i,0})$ for $i = 1, \ldots, r$ we can extend our system (2.30) to

$$F^{(1)} := \begin{bmatrix} F_{n-1}^{(1)} \\ \vdots \\ F_0^{(1)} \end{bmatrix} \in \hat{A}[\boldsymbol{G}_r, \ldots, \boldsymbol{G}_1]^n \tag{2.34}$$

where (again with $G_{i,n_i} = 1$)

$$F_k^{(1)} := \sum_{k_1 + \ldots + k_r = k} G_{1,k_1} \cdots G_{r,k_r} - f_k\ , \quad 0 \leq k < n\ .$$

Define $\boldsymbol{g}_i := (g_{n_i-1}, \ldots, g_0) \in \hat{A}^{n_i}$ then

$$\left\| F^{(1)}(\boldsymbol{g}_r, \ldots, \boldsymbol{g}_1) \right\|_{\mathfrak{m}} < 1$$

Hence we can uniquely lift the $g_1, \ldots, g_r$ if the Jacobian $J_{F^{(1)}}(\boldsymbol{g}_r, \ldots, \boldsymbol{g}_1)$ is invertible. Therefore now we take a closer look at the Jacobian of $F^{(1)}$.

Lemma 2.59. Let $F^{(1)} \in \hat{A}[\boldsymbol{G}_r, \ldots, \boldsymbol{G}_1]^n$ be the system (2.34) and set

$$G_i := \sum_{j=0}^{n_i-1} G_{i,j} X^j + X^{n_i} \in \hat{A}[\boldsymbol{G}_i][X] \;, \quad i = 1, \ldots, r \;.$$

Then

$$\det\left(J_{F^{(1)}}\right) = \delta \prod_{1 \le i < j \le r} \operatorname{res}(G_i, G_j)$$

where $\delta \in \{-1, 1\}$.

Proof. First we show that $\operatorname{res}(G_i, G_j)$ divides $\det\left(J_{F^{(1)}}\right)$ for $1 \le i < j \le r$. The idea is to construct maps $\varphi^{(i,j)}$ and $\psi^{(i,j)}$ with $\det\left(J_{\psi^{(i,j)}}\right) = \operatorname{res}(G_i, G_j)$ such that $F^{(1)} = \varphi^{(i,j)} \circ \psi^{(i,j)}$ and then to make use of the chain rule.

We fix $1 \le i < j \le r$ and without loss of generality we assume $i = 1$ and $j = 2$. For the product $G_1 G_2$ we can construct analogous to (2.30) the system

$$H^{(1,2)} := \begin{bmatrix} H^{(i,j)}_{n_2+n_1-1} \\ \vdots \\ H^{(i,j)}_0 \end{bmatrix} \in \hat{A}[\boldsymbol{G}_2, \boldsymbol{G}_1]^{n_2+n_1} \;.$$

where (with $G_{1,n_1} = G_{2,n_2} = 1$)

$$H^{(1,2)}_k := \sum_{k_1+k_2=k} G_{1,k_1} G_{2,k_2} \in \hat{A}[\boldsymbol{G}_2, \boldsymbol{G}_1] \;, \quad 0 \le k < n_2 + n_1 \;.$$

Notice that by construction

$$J_{H^{(1,2)}} = \operatorname{Syl}(G_2, G_1) \;. \tag{2.35}$$

Moreover, consider the maps

$$\begin{array}{rccc} \psi^{(1,2)} \;: & \hat{A}^n & \to & \hat{A}^{n-n_1-n_2} \times \hat{A}^{n_2+n_1}, \\ & (\boldsymbol{G}_3, \ldots, \boldsymbol{G}_r, \boldsymbol{G}_2, \boldsymbol{G}_1) & \mapsto & (\boldsymbol{G}_3, \ldots, \boldsymbol{G}_r, H^{(1,2)}(\boldsymbol{G}_2, \boldsymbol{G}_1)) \end{array}$$

and

$$\begin{array}{rccc} \varphi^{(1,2)} \;: & \hat{A}^{n-n_1-n_2} \times \hat{A}^{n_2+n_1} & \to & \hat{A}^n, \\ & ((\boldsymbol{G}_3, \ldots, \boldsymbol{G}_r), \boldsymbol{H}) & \mapsto & \begin{bmatrix} \varphi^{(1,2)}_{n-1}(\boldsymbol{G}_3, \ldots, \boldsymbol{G}_r, \boldsymbol{H}) \\ \vdots \\ \varphi^{(1,2)}_0(\boldsymbol{G}_3, \ldots, \boldsymbol{G}_r, \boldsymbol{H}) \end{bmatrix} \end{array}$$

where for $0 \le k < n$

$$\begin{array}{rccc} \varphi^{(1,2)}_k \;: & \hat{A}^{n-n_1-n_2} \times \hat{A}^{n_2+n_1} & \to & \hat{A}, \\ & ((\boldsymbol{G}_3, \ldots, \boldsymbol{G}_r), \boldsymbol{H}) & \mapsto & \displaystyle\sum_{k_3+\cdots+k_r+\tau=k} G_{3,k_3} \cdots G_{r,k_r} H_\tau - f_k \end{array}$$

with $\boldsymbol{H} = (H_{n_2+n_1-1}, \dots, H_0)$ and $H_{n_2+n_1} = 1$. Then

$$F^{(1)}(\boldsymbol{G}_r, \dots, \boldsymbol{G}_1) = \varphi^{(1,2)} \circ \psi^{(1,2)}(\boldsymbol{G}_3, \dots, \boldsymbol{G}_r, \boldsymbol{G}_2, \boldsymbol{G}_1)$$

and by the chain rule it follows with $\delta \in \{-1, 1\}$

$$\det\left(J_{F^{(1)}}\right) = \delta \det\left(J_{\varphi^{(1,2)} \circ \psi^{(1,2)}}\right) = \delta \det\left(\left(J_{\varphi^{(1,2)}} \circ \psi^{(1,2)}\right) J_{\psi^{(1,2)}}\right) .$$

Finally we obtain

$$\begin{aligned}\det\left(J_{\psi^{(1,2)}}\right) &= \det\left(\begin{bmatrix} I_{n-n_1-n_2} & \\ & J_{H^{(1,2)}} \end{bmatrix}\right) = \det\left(J_{H^{(1,2)}}\right) \\ &\overset{(2.35)}{=} \det(\mathrm{Syl}(G_2, G_1)) = \mathrm{res}(G_2, G_1) .\end{aligned}$$

Therefore $\mathrm{res}(G_i, G_j)$ divides $\det(J_{F^{(1)}})$ for all $1 \le i < j \le r$.

From $\mathrm{Syl}(G_i, G_j)$ it follows easily that $\deg(\mathrm{res}(G_i, G_j)) = n_i + n_j - 1$ (remember that G_i and G_j are monic). Since the $\mathrm{res}(G_i, G_j)$ are clearly pairwise relatively prime we obtain

$$\deg(\det(J_{F^{(1)}})) \ge \sum_{1 \le i < j \le r} (n_i + n_j - 1) = (r-1)n - \sum_{1 \le i < j \le r} 1 = (r-1)n - \frac{1}{2}(r-1)r .$$

On the other hand, for $k = 1, \dots, r$, the term with highest degree in $F^{(1)}_{n-k}$ is $\prod_{s \in S} G_{s,n_s-1}$ for a subset $S \subset \{1, \dots, r\}$ with $|S| = k$ (again the G_i are monic). For $k > r$ the degree of $F^{(1)}_{n-k}$ is at most r. Hence

$$\begin{aligned}\deg(\det(J_{F^{(1)}})) &\le \sum_{k=1}^{r} (k-1) + (n-r)(r-1) = \sum_{k=1}^{r-1} k + (n-r)(r-1) \\ &= \frac{1}{2}(r-1)r + n(r-1) - r(r-1) \\ &= (r-1)n - \frac{1}{2}(r-1)r\end{aligned}$$

and we conclude $\det(J_{F^{(1)}}) = \delta \prod_{1 \le i < j \le r} \mathrm{res}(G_i, G_j)$. □

With the help of this lemma we can state the first generalization of the Hensel lifting theorem.

Proposition 2.60. Let $\hat{A}$ be a complete Noetherian local ring with maximal ideal $\mathfrak{m}$ and $f \in \hat{A}[X]$ a non-zero polynomial such that $\mathrm{lc}(f) \notin \mathfrak{m}$. Let $g_1, \dots, g_r \in \hat{A}[X]$ be pairwise strongly relatively prime polynomials such that

$$f \equiv \prod_{i=1}^{r} g_i \bmod \mathfrak{m} .$$

Then there exists unique pairwise strongly relatively prime polynomials $\hat{g}_i \in \hat{A}[X]$, $i = 1, \dots, r$, such that

$$f = \prod_{i=1}^{r} \hat{g}_i \quad \text{and} \quad \hat{g}_i \equiv g_i \bmod \mathfrak{m} , \quad 1 \le i \le r . \tag{2.36}$$

Proof. With the system (2.34) the proof is largely analog to the previous proof of our Hensel lifting theorem 2.56. Assume again without loss of generality that f and $g_1, \dots, g_r$ are monic. Considering the previous notation the only thing that remains to show is that $\det(J_{F^{(1)}}(\boldsymbol{g}_r, \dots, \boldsymbol{g}_1))$ is a unit and that the lifted polynomials are pairwise strongly relatively prime.

Since the g_i are pairwise strongly relatively prime $\operatorname{res}(g_i, g_j)$ is a unit for $1 \leq i < j \leq r$ by theorem 2.47 and thus $\det(J_{F^{(1)}}(\boldsymbol{g}_r, \dots, \boldsymbol{g}_1))$ is a unit by lemma 2.59. Therefore by corollary 2.43 to the Newton iteration we can obtain a unique $(\hat{\boldsymbol{g}}_r, \dots, \hat{\boldsymbol{g}}_1) \in \hat{A}^{n_r+\dots+n_1}$ such that

$$F^{(1)}(\hat{\boldsymbol{g}}_r, \dots, \hat{\boldsymbol{g}}_1) = \boldsymbol{0},\ \|(\hat{\boldsymbol{g}}_r, \dots, \hat{\boldsymbol{g}}_1) - (\boldsymbol{g}_r, \dots, \boldsymbol{g}_1)\|_{\mathfrak{m}} < 1 \text{ and } \det(J_F^{(1)}(\hat{\boldsymbol{g}}_r, \dots, \hat{\boldsymbol{g}}_1)) \text{ unit.}$$

If we write $\hat{\boldsymbol{g}}_i = (\hat{g}_{i,n_i-1}, \dots, \hat{g}_{i,0})$ for $i = 1, \dots, r$ we can define the unique polynomials

$$\hat{g}_i := \sum_{j=0}^{n_i-1} \hat{g}_{i,j} X^j + X^{n_i} \in \hat{A}[X]\ ,\ 1 \leq i \leq r$$

which satisfy (2.36). Furthermore, by lemma 2.59 it follows that

$$\delta \prod_{1 \leq i < j \leq r} \operatorname{res}(\hat{g}_i, \hat{g}_j) = \det\left(J_{F^{(1)}}(\hat{\boldsymbol{g}}_r, \dots, \hat{\boldsymbol{g}}_1)\right)\ ,\ \delta \in \{-1, 1\}$$

is a unit. Therefore $\hat{g_1}, \dots, \hat{g_r}$ are pairwise strongly relatively prime by theorem 2.47 □

Now we are ready to handle the general case. Unless otherwise stated we continue the notation of the previous case. Suppose we have pairwise strongly relatively prime monic polynomials $g_1, \dots, g_r \in \hat{A}[X]$ with $m_i := \deg(g_i)$, $m := \sum_{i=1}^r m_i$ and positive integers $e_1, \dots, e_r$ such that

$$f \equiv \prod g_i^{e_i} \bmod \mathfrak{m}$$

and assume that there exists an index i with $e_i \geq 2$. To simplify the notation define the polynomials

$$H_i := \Big(\sum_{j=0}^{m_i-1} G_{i,j} X^j + X^{m_i} \Big)^{e_i} \in \hat{A}[\boldsymbol{G}_r, \dots, \boldsymbol{G}_1][X]$$

and write $H_i = \sum_{k=0}^{n_i-1} H_{i,k} X^k + X^{n_i}$ with $n_i := m_i e_i$. Then (again with $G_{i,m_i} = 1$)

$$H_{i,k} = \sum_{k_1+\dots+k_{e_i}=k} G_{i,k_1} \dots G_{i,k_{e_i}} \in \hat{A}[\boldsymbol{G}_r, \dots, \boldsymbol{G}_1]\ ,\ 0 \leq k < n_i$$

for $i = 1, \dots, r$. Now we can extend our previous system (2.34) to

$$F^{(2)} := \begin{bmatrix} F_{n-1}^{(2)} \\ \vdots \\ F_0^{(2)} \end{bmatrix} \in \hat{A}[\boldsymbol{G}_r, \dots, \boldsymbol{G}_1]^n \tag{2.37}$$

where (again with $H_{i,n_i} = 1$)

$$F_k^{(2)} := \sum_{k_1+\ldots+k_r=k} H_{1,k_1} H_{2,k_2} \cdots H_{r,k_r} - f_k \quad , \quad 0 \leq k < n \; .$$

But in contrast to the previous case the system $F^{(2)}$ is overdetermined since

$$\deg(f) = n = \sum_{i=1}^{r} n_i = \sum_{i=1}^{r} m_i e_i > \sum_{i=1}^{r} m_i \; .$$

Therefore we have to show that there exists a consistent subsystem of $F^{(2)}$ such that the Jacobian of this subsystem is invertible for $(\boldsymbol{g}_r, \ldots, \boldsymbol{g}_1)$.

In preparation of the proof we take a closer look at $F^{(2)}$. Consider the map

$$\varphi : \hat{A}^m \to \hat{A}^n, \quad \begin{bmatrix} \boldsymbol{g}_r \\ \vdots \\ \boldsymbol{g}_1 \end{bmatrix} \mapsto \begin{bmatrix} \varphi^{(r)}(\boldsymbol{g}_r) \\ \vdots \\ \varphi^{(1)}(\boldsymbol{g}_1) \end{bmatrix}$$

where

$$\varphi^{(i)} : \hat{A}^{m_i} \to \hat{A}^{n_i}, \; \boldsymbol{g} \mapsto \begin{bmatrix} H_{i,n_i-1}(\boldsymbol{g}) \\ \vdots \\ H_{i,0}(\boldsymbol{g}) \end{bmatrix}$$

for $1 \leq i \leq r$. Then

$$F^{(2)} = F^{(1)} \circ \varphi$$

and by the chain rule

$$J_{F^{(2)}} = (J_{F^{(1)}} \circ \varphi)\, J_\varphi \; . \tag{2.38}$$

To determine a consistent subsystem of $F^{(2)}$ it is therefore sufficient to consider $F^{(1)} \circ \varphi$ and $(J_{F^{(1)}} \circ \varphi)\, J_\varphi$ respectively.

Theorem 2.61 (Generalized Hensel lifting)**.** Let $\hat{A}$ be a complete Noetherian local ring with maximal ideal $\mathfrak{m}$ and $f \in \hat{A}[X]$ a non-zero polynomial such that $\mathrm{lc}(f) \notin \mathfrak{m}$. Let $g_1, \ldots, g_r \in \hat{A}[X]$ be pairwise strongly relatively prime polynomials and $e_1, \ldots, e_r$ positive integers such that

$$f \equiv \prod_{i=1}^{r} g_i^{e_i} \bmod \mathfrak{m}$$

and $e_i \not\equiv 0 \bmod \mathfrak{m}$, $i = 1, \ldots, r$. Then there exists unique pairwise strongly relatively prime polynomials $\hat{g}_1, \ldots, \hat{g}_r \in \hat{A}[X]$ such that

$$f = \prod_{i=1}^{r} \hat{g}_i^{e_i} \quad \text{and} \quad \hat{g}_i \equiv g_i \bmod \mathfrak{m} \; , \; 1 \leq i \leq r \; .$$

Proof. We make again use of the previous introduced notation and assume again that, without loss of generality, f and $g_1 \ldots, g_r$ are monic. If we can show that the system (2.37) has a consistent subsystem G such that $\det(J_G(\boldsymbol{g}_r, \ldots, \boldsymbol{g}_1))$ is a unit, the proof is largely identical to the proof of the previous corollary. The only thing that remains to show is that the lifted polynomials are pairwise strongly relatively prime and unique.

For $i = 1, \dots, r$ set $\boldsymbol{h}_i := (h_{i,n_i-1}, \dots, h_{i,0}) := \varphi^{(i)}(\boldsymbol{g_i})$ and define the polynomials $h_i := \sum_{k=0}^{n_i-1} h_{i,k} X^k + X^{n_i}$. Then we have $h_i = g_i^{e_i}$ and by (2.38)

$$\begin{aligned} J_{F^{(2)}}(\boldsymbol{g}_r, \dots, \boldsymbol{g}_1) &= J_{F^{(1)} \circ \varphi}(\boldsymbol{g}_r, \dots, \boldsymbol{g}_1) \\ &= J_{F^{(1)}}(\boldsymbol{h}_r, \dots, \boldsymbol{h}_1) J_{\varphi}(\boldsymbol{g}_r, \dots, \boldsymbol{g}_1) \, . \end{aligned}$$

Since the g_i are strongly relatively prime the h_i are also strongly relatively prime. Thus it follows that

$$\det(J_{F^{(1)}}(\boldsymbol{h}_r, \dots, \boldsymbol{h}_1)) = \delta \prod_{1 \le i < j \le r} \operatorname{res}(h_i, h_j) \quad , \quad \delta \in \{-1, 1\}$$

is a unit by theorem 2.47 and lemma 2.59 and in particular that $J_{F^{(1)}}(\boldsymbol{h}_r, \dots, \boldsymbol{h}_1)$ is invertible. Now by construction

$$J_{\varphi}(\boldsymbol{g}_r, \dots, \boldsymbol{g}_1) = \begin{bmatrix} J_{\varphi^{(r)}}(\boldsymbol{g}_r) & & \\ & \ddots & \\ & & J_{\varphi^{(1)}}(\boldsymbol{g}_1) \end{bmatrix}$$

where $J_{\varphi^{(i)}}(\boldsymbol{g}_r) \in \hat{A}^{n_i, m_i}$ for $i = 1, \dots, r$. Since the g_i are assumed to be monic we have

$$\frac{\partial H_{i,n_i-k}}{\partial G_{i,m_i-j}} = \begin{cases} e_i & , \ k = j \\ 0 & , \ k < j \\ * & , \ \text{else} \end{cases}$$

for $1 \le i \le r$, $1 \le k \le n_i$ and $1 \le j \le m_i$. Therefore it follows that

$$J_{\varphi^{(i)}} = \begin{bmatrix} e_i & & \\ * & \ddots & \\ \vdots & \ddots & e_i \\ \vdots & \ddots & * \\ \vdots & \ddots & \vdots \\ * & \dots & * \end{bmatrix} \, .$$

Consider $J_{F^{(1)}}(\boldsymbol{h}_r, \dots, \boldsymbol{h}_1) J_{\varphi}(\boldsymbol{g}_r, \dots, \boldsymbol{g}_1)$ as an element in $(\hat{A}/\mathfrak{m})^{n,m}$. Due to the fact that $\hat{A}/\mathfrak{m}$ is a field we can apply vector space theory:

Since $J_{F^{(1)}}(\boldsymbol{h}_r, \dots, \boldsymbol{h}_1)$ is invertible it has full rank and by our previous observation $\operatorname{rank}(J_{\varphi}(\boldsymbol{g}_r, \dots, \boldsymbol{g}_1)) = m$. Thus

$$\operatorname{rank}(J_{F^{(2)}}(\boldsymbol{g}_r, \dots, \boldsymbol{g}_1)) \overset{(2.38)}{=} \operatorname{rank}(J_{F^{(1)}}(\boldsymbol{h}_r, \dots, \boldsymbol{h}_1) J_{\varphi}(\boldsymbol{g}_r, \dots, \boldsymbol{g}_1)) = m$$

and we can determine a consistent subsystem G with $\det(J_G)(\boldsymbol{g}_r, \dots, \boldsymbol{g}_1) \notin \mathfrak{m}$. This also implies that $\det(J_G)(\boldsymbol{g}_r, \dots, \boldsymbol{g}_1))$ is a unit in $\hat{A}$.

Hence we can analogous to the previous proof apply the Newton iteration on the system G to determine polynomials $\hat{g}_1, \dots, \hat{g}_r \in \hat{A}[X]$ such that $f = \prod_{i=1}^{r} \hat{g}_i^{e_i}$ and $\hat{g}_i \equiv g_i \bmod \mathfrak{m}$ for $i = 1, \dots, r$. If we define $\hat{h}_i := \hat{g}^{e_i} = \sum_{k=0}^{n_i-1} \hat{h}_{i,k} X^k + X^{n_i}$ and $\hat{\boldsymbol{h}}_i := (\hat{h}_{i,n_i-1}, \dots, \hat{h}_{i,0})$

for $i = 1, \dots, r$ we have that

$$\det(J_{F^{(1)}}(\hat{\boldsymbol{h}}_r, \dots, \hat{\boldsymbol{h}}_1)) = \delta \prod_{1 \le i < j \le r} \operatorname{res}(\hat{h}_i, \hat{h}_j) \ , \ \delta \in \{-1, 1\}$$

is a unit. Therefore $\hat{h}_r, \dots, \hat{h}_1$ are pairwise strongly relatively prime by theorem 2.47 and thus the $\hat{g}_i$ are also pairwise strongly relatively prime.

Finally we have to show that the $\hat{g}_r, \dots, \hat{g}_1$ are unique. As $\hat{g}_i \equiv g_i \bmod \hat{\mathfrak{m}}$ for $i = 1, \dots, r$ we also have $\hat{h}_i \equiv h_i \bmod \mathfrak{m}$ for $i = 1, \dots, r$. Since also $f = \prod_{i=1}^r \hat{h}_i$ and $f \equiv \prod_{i=1}^r h_i \bmod \mathfrak{m}$ it follows that $\hat{h}_r, \dots, \hat{h}_1$ are unique by our previous proposition 2.60 and thus $\hat{g}_r, \dots, \hat{g}_1$ are also unique. □

We have seen that we can uniquely lift a factorization in a complete Noetherican local ring $\hat{A}$. A downside is that since each element in $\hat{A}$ is a possibly infinite series it is not possible to compute exactly in $\hat{A}$. But the situation can be rescued.

Corollary 2.62. Let A be a Noetherian ring with maximal ideal $\mathfrak{m}$ and $f \in A[X]$ a non-zero polynomial such that $\operatorname{lc}(f) \notin \mathfrak{m}$. Let $g_1, \dots, g_r \in A[X]$ be polynomials and $e_1, \dots, e_r$ positive integers such that

$$f \equiv \prod_{i=1}^r g_i^{e_i} \bmod \mathfrak{m} \quad \text{and} \quad \gcd(g_i \bmod \mathfrak{m}, g_j \bmod \mathfrak{m}) \notin \mathfrak{m} \text{ for } 1 \le i < j \le r$$

and $e_i \not\equiv 0 \bmod \mathfrak{m}$, $i = 1, \dots, r$. Then there exists for all positive integers D polynomials $g_1^{(D)}, \dots, g_r^{(D)} \in A[X]$ such that the $g_i^{(D)}$ are pairwise strongly relatively prime in $A/\mathfrak{m}^D$,

$$f \equiv \prod_{i=1}^r \left(g_i^{(D)}\right)^{e_i} \bmod \mathfrak{m}^D \quad \text{and} \quad g_i^{(D)} \equiv g \bmod \mathfrak{m} \text{ for } 1 \le i \le r. \tag{2.39}$$

Furthermore, the $g_1^{(D)}, \dots, g_r^{(D)}$ are unique, i.e., for all $g_1^*, \dots, g_r^* \in A[X]$ with $f \equiv \prod_{i=1}^r (g_i^*)^{e_i} \bmod \mathfrak{m}^D$ and $g_i^* \equiv g \bmod \mathfrak{m}$, for $1 \le i \le r$, we have

$$g_i^* - g_i^{(D)} \equiv 0 \bmod \mathfrak{m}^D \ , \ i = 1, \dots, r \ . \tag{2.40}$$

Moreover, there exists unique strongly relatively prime polynomials $\hat{g}_1, \dots, \hat{g}_r \in \hat{A}_{\mathfrak{m}}[X]$ such that

$$f = \prod_{i=1}^r \hat{g}_i^{e_i} \quad \text{and} \quad \hat{g}_i \equiv g_i \bmod \hat{\mathfrak{m}} \text{ for } 1 \le i \le r$$

where $\hat{A}_{\mathfrak{m}}$ is the completion of A with respect to $\mathfrak{m}$ with maximal ideal $\hat{\mathfrak{m}}$.

Proof. By corollary 2.55 the g_i are pairwise strongly relatively prime in $\hat{A}_{\mathfrak{m}}$. Therefore by theorem 2.61 there exists unique strongly relatively prime polynomials

$$\hat{g}_i = \sum_{j=0}^{n_i} \hat{g}_j^{(i)} X^j \in \hat{A}_{\mathfrak{m}}[X] \ , \ 1 \le i \le r,$$

such that $f = \prod_{i=1}^r \hat{g}_i^{e_i}$ and $\hat{g}_i \equiv g_i \bmod \hat{\mathfrak{m}}$ for $1 \le i \le r$. If we write for $i = 1, \dots, r$

$$\hat{g}_j^{(i)} = (\hat{g}_{j,1}^{(i)} + \mathfrak{m}, \hat{g}_{j,2}^{(i)} + \mathfrak{m}^2, \dots) \ , \ 0 \le j \le n_i \ ,$$

then the statements (2.39) and (2.40) are satisfied by the polynomials

$$g_i^{(D)} := \hat{g}_{n_i,D}^{(i)} X^{n_i-1} + \ldots + \hat{g}_{0,D}^{(i)} \in A[X] \ , \ 1 \leq i \leq r$$

for all positive integers D. The $g_i^{(D)}$ are unique and pairwise strongly relatively prime due to the fact that the $\hat{g}_i$ are unique and pairwise strongly relatively prime. □

Remark 2.63. The proof still relies on computations in $\hat{A}_{\mathfrak{m}}$ to compute in each iteration step the inverse of the Jacobian $J_G\left(\boldsymbol{g}^{k-1}\right)$ of the consistent subsystem G from the proof of theorem 2.61. But we already showed in our Newton iteration algorithm that the inversion of $J_G\left(\boldsymbol{g}^{k-1}\right)$ can be replaced by the computation of $J^{(k)} \in \hat{A}_{\mathfrak{m}}^{m,m}$ such that $\|J^{(k)} J_G\left(\boldsymbol{g}^{k-1}\right) - I_m\|_{\hat{\mathfrak{m}}} \leq 2^{-2^k}$ and this in fact nothing else than the inversion of $J_G\left(\boldsymbol{g}^{k-1}\right)$ in $A/\mathfrak{m}^{2^k}$!

The proof of the previous corollary is constructive and together with remark 2.63 yields the following Hensel lifting algorithm. But before we start we need to introduce some additional notation. For positive integers n denote by $[n]$ the set $\{1, \ldots, n\}$ and for $C = (\boldsymbol{c}_{i,.})_{1 \leq i \leq n} \in A^{n,m}$ and an index set $I \subset [n]$ denote by C_I the matrix $(\boldsymbol{c}_{i,.})_{i \in I} \in A^{|I|,m}$. Moreover, for an ideal $J \subset A$ denote by $C \bmod J$ the reduction of each element in C modulo J.

Algorithm 2.64 Hensel lifting

Input: Noetherian ring A with maximal ideal $\mathfrak{m}$, $D \in \mathbb{N}_{>0}$ and $f, g_1, \ldots, g_r \in A[X]$ and integers $e_1, \ldots, e_r$ such that $\mathrm{lc}(f) \not\equiv 0 \bmod \mathfrak{m}$, $e_i \not\equiv 0 \bmod \mathfrak{m}$, $f \equiv \prod_{i=1}^r g_i^{e_i} \bmod \mathfrak{m}$ and $\gcd(g_i \bmod \mathfrak{m}, g_j \bmod \mathfrak{m}) \notin \mathfrak{m}$ for $1 \leq i < j \leq r$.

Output: $\hat{g}_1, \ldots, \hat{g}_r \in A[X]$ such that $f \equiv \prod_{i=1}^r \hat{g}_i^{e_i} \bmod \mathfrak{m}^D$. Furthermore the $\hat{g}_i$ are unique and pairwise strongly relatively prime in $A/\mathfrak{m}^D$.

// Normalize input polynomials
Compute *lead_f_inv* such that $lead\_f\_inv \cdot \mathrm{lc}(f) \equiv 1 \bmod \mathfrak{m}^D$
$\tilde{f} := lead\_f\_inv \cdot f$
$n := \deg(f)$
for $i := 1, \ldots, r$ **do**
    // save coefficients for output polynomials
    $lead\_g_i := \mathrm{lc}(g_i) \bmod \mathfrak{m}$
    Compute $lead\_g_i\_inv$ such that $lead\_g_i\_inv \cdot lead\_g_i \equiv 1 \bmod \mathfrak{m}$
    $g_i^{(0)} := lead\_g_i\_inv \cdot g_i \bmod \mathfrak{m}$
    $n_i := \deg(g_i^{(0)})$
end for
$m := \sum_{i=1}^r n_i$
// Create coefficient arrays
for $i := 1, \ldots, r$ **do**
    $\boldsymbol{g}_i^{(0)} := [g_{i,n_i-1}^{(0)}, \ldots, g_{i,0}^{(0)}]$
end for
/* *next page* */

// Preparation for Newton iteration
Create $F^{(2)} \in A[\boldsymbol{G}_r, \ldots, \boldsymbol{G}_1]^n$ as defined in (2.37)
Create the corresponding Jacobian $J_{F^{(2)}} \in A[\boldsymbol{G}_r, \ldots, \boldsymbol{G}_1]^{n,m}$
/* *continue Hensel lifting algorithm* */
if $m \neq n$ **then** $//F^{(2)}$ overdetermined
  Determine $I \subset [n]$, $|I| = m$, s.th. rank $\left((J_{F^{(2)}})_I \left(\boldsymbol{g}_r^{(0)}, \ldots, \boldsymbol{g}_1^{(0)} \right) \right) = m$ in $A/\mathfrak{m}$
  $G := F_I^{(2)}$
  $J_G := (J_{F^{(2)}})_I$
else
  $G := F^{(2)}$
  $J_G := J_{F^{(2)}}$
end if
// Newton Iteration
$d := \lceil \log_2 D \rceil$
for $k := 1, \ldots, d$ **do**
  Compute $J^{(k)} \in A^{m,m}$ such that $J^{(k)} J_G(\boldsymbol{g}_r^{(k-1)}, \ldots, \boldsymbol{g}_1^{(k-1)}) \equiv I_m \bmod \mathfrak{m}^{2^k}$
  $\left(\boldsymbol{g}_r^{(k)}, \ldots, \boldsymbol{g}_1^{(k)} \right) := \left(\boldsymbol{g}_r^{(k-1)}, \ldots, \boldsymbol{g}_1^{(k-1)} \right) - J^{(k)} G\left(\boldsymbol{g}_r^{(k-1)}, \ldots, \boldsymbol{g}_1^{(k-1)} \right) \bmod \mathfrak{m}^{2^k}$
end for
// Create output polynomials
for $i := 1, \ldots, r$ **do**
  $\hat{g}_i := \sum_{k=0}^{n_i - 1} lead\_g_i \cdot g_{i,k}^{(d)} X^k + lead\_g_i X^{n_i}$
end for
return $\hat{g}_1, \ldots, \hat{g}_r$

Theorem 2.65. The Hensel lifting algorithm 2.64 works correctly and needs at most $\mathcal{O}((\log_2(D) + 1) \deg(f)^3)$ arithmetic operations.

Proof. The correctness of the algorithm follows from corollary 2.62, remark 2.63, the proof of theorem 2.61 and the correctness of the Newton iteration algorithm 2.41. Since the dominant step is the Newton iteration it follows that the algorithm needs at most $\mathcal{O}((\log_2(D) + 1) \deg(f)^3)$ arithmetic operations. □

Chapter 3

Evaluations of multivariate polynomials

Before we start with this chapter, we have to fix some notation. We denote for a finite set S by $\mathrm{card}(S)$ the cardinality of S. For a field k consider the polynomial $f \in k[X_1, \ldots, X_n]$. We denote by $\deg(f)$ the *total* degree of f and for a family $I \subset \{X_1, \ldots, X_n\}$ of indeterminates we denote by $\deg_I(f)$ the (total) degree of f with respect to I. Moreover, we denote for an indeterminate X_i by $\mathrm{lc}_{X_i}(f)$ the leading coefficient of f with respect to X_i. Similarly we denote for polynomials $g, h \in k[X_1, \ldots, X_n, Y]$ by $\mathrm{res}_Y(g, h)$ and $\mathrm{Syl}_Y(g, h)$ the resultant / Sylvester matrix of g and h where g and h are considered as polynomials in $k[X_1, \ldots, X_n][Y]$.

Example 3.1. Let $f = 3X^2Y^2Z^3 - 2XY^4Z \in \mathbb{Q}[X, Y, Z]$. Then

$$\deg(f) = 7,\ \deg_Y(f) = 4 \text{ and } \mathrm{lc}_Y(f) = -2XZ\ .$$

3.1 Effective Hilbert irreducibility

Our goal in this section is to derive an effective version of Hilbert's Irreducibility theorem. We will follow the original publication from Kaltofen [10]. We show that a certain bivariate image of an irreducible polynomial $f \in k[X_1, \ldots, X_n]$ remains irreducible with a controllable high probability.

We start with the fundamental Schwartz-Zippel lemma. This lemma will be used in nearly every proof of this chapter.

Lemma 3.2 (Schwartz-Zippel lemma)**.** Let A be an integral domain, f a non-zero polynomial in $A[X_1, \ldots, X_n]$ with total degree D. Let $S \subset A$ be a finite subset. Then the probability

$$\mathrm{Prob}\ \big(\ f(a_1, \ldots, a_n) = 0 \,|\, a_1, \ldots, a_n \in S\ \big) \leq \frac{D}{\mathrm{card}(S)}\ .$$

Proof. We prove the lemma by induction over the number of indeterminates. Let $n = 1$ and $f \in A[X_1]$. Since an univariate polynomial with degree D has at most D roots it

follows Prob $\left(f(a_1) = 0 \,|\, a_1 \in S \right) \leq D/\mathrm{card}(S)$.
Now assume the hypothesis holds for all polynomials in $A[X_1, \ldots, X_{n-1}]$ and let f be a polynomial in $A[X_1, \ldots, X_n]$. Set $d = \deg_{X_n}(f)$ and $f_d = \mathrm{lc}_{X_n}(f) \in A[X_1, \ldots, X_{n-1}]$. Then $\deg(f_d) \leq D - d$ and by our induction hypothesis

$$\text{Prob } \left(f_d(a_1, \ldots, a_{n-1}) = 0 \,|\, a_1, \ldots, a_{n-1} \in S \right) \leq \frac{D-d}{\mathrm{card}(S)} .$$

If we have $f_d(a_1, \ldots, a_{n-1}) \neq 0$ for $a_1, \ldots, a_{n-1} \in S$ it follows that $\deg f(a_1, \ldots, a_{n-1}, X_n) = d$ and thus that there are at most d roots of $f(a_1, \ldots, a_{n-1}, X_n)$ in S. Hence, by our induction hypothesis

$$\text{Prob}(\ f(a_1, \ldots, a_n) = 0 \,|\, f_d(a_1, \ldots, a_{n-1}) \neq 0, \ a_n \in S \ ) \leq \frac{d}{\mathrm{card}(S)}.$$

We can now conclude that for arbitrary $a_1, \ldots, a_n \in S$ with $\boldsymbol{a} := (a_1, \ldots, a_n)$

$$\begin{aligned}
\text{Prob } \left(f(\boldsymbol{a}) = 0 \right) &= \text{Prob } \left(f(\boldsymbol{a}) = 0 \,|\, f_d(a_1, \ldots, a_{n-1}) = 0 \right) \text{Prob } \left(f_d(a_1, \ldots, a_{n-1}) = 0 \right) \\
&\quad + \text{Prob } \left(f(\boldsymbol{a}) = 0 \,|\, f_d(a_1, \ldots, a_{n-1}) \neq 0 \right) \text{Prob } \left(f_d(a_1, \ldots, a_{n-1}) \neq 0 \right) \\
&\leq \text{Prob } \left(f_d(a_1, \ldots, a_{n-1}) = 0 \right) + \text{Prob } \left(f(\boldsymbol{a}) = 0 \,|\, f_d(a_1, \ldots, a_{n-1}) \neq 0 \right) \\
&\leq \frac{D-d}{\mathrm{card}(S)} + \frac{d}{\mathrm{card}(S)} \\
&= \frac{D}{\mathrm{card}(S)}
\end{aligned}$$

□

Remark 3.3. Note that the probability only depends on the degree of f and on the cardinality of S and not on the number of indeterminates!

Lemma 3.4 ([10])**.** Let k be a field and $f \in k[X_1, \ldots, X_n, Y]$ an irreducible polynomial with $\partial f/\partial Y \neq 0$, $d = \deg_Y(f)$ and $D = \deg_{X_1,\ldots,X_n}(f)$. Pick random elements $a_1, \ldots, a_n$ from a finite subset $S \subset k$. Then

$$\text{Prob } \left(f(a_1, \ldots, a_n, Y) \text{ square free} \wedge \mathrm{lc}_Y(f)(a_1, \ldots, a_n) \neq 0 \right) \geq 1 - \frac{(2d+1)D}{\mathrm{card}(S)} .$$

Proof. Since f is irreducible and $\partial f/\partial Y \neq 0$ we have $\gcd(f, \partial f/\partial Y) = 1$. Therefore the resultant

$$r_f(X_1, \ldots, X_n) := \mathrm{res}_Y \left(f, \frac{\partial f}{\partial Y} \right) = \det \left(\mathrm{Syl}_Y \left(f, \frac{\partial f}{\partial Y} \right) \right) \neq 0$$

by corollary 2.51. Since every term of r_f is a product of $d+(d-1)$ coefficients f_i of f with degree at most D it follows $\deg(r_f) \leq (2d-1)D$. We write $\partial f/\partial Y = k f_k Y^{k-1} + \ldots + f_1$ with $f_i \in k[X_1, \ldots, X_n]$, $k f_k \neq 0$ and $\deg(f_i) \leq D$, $1 \leq i \leq k$.

We claim that if we select elements $a_1, \ldots, a_n \in S$ such that $(\mathrm{lc}_Y(f) k f_k r_f)(a_1, \ldots, a_n) \neq 0$ then $\bar{f}(Y) := f(a_1, \ldots, a_n, Y)$ is square free. Assume this were not the case. Then $\gcd(\bar{f}, \partial \bar{f}/\partial Y) \neq 1$ and thus $\mathrm{res}_Y(\bar{f}, \partial \bar{f}/\partial Y) = 0$ by corollary 2.51. But $\mathrm{res}_Y(\bar{f}, \partial \bar{f}/\partial Y) = r_f(a_1, \ldots, a_n) \neq 0$, a contradiction. Hence $\bar{f}$ is square free if $(\mathrm{lc}_Y(f) k f_k r_f)(a_1, \ldots, a_n) \neq 0$.

Since $\deg(\mathrm{lc}_Y(f)kf_kr_f) \leq D + D + (2d-1)D = (2d+1)D$ it follows that

$$\mathrm{Prob}\left((\mathrm{lc}_Y(f)kf_kr_f)(a_1,\ldots,a_n) \neq 0 \,|\, a_1,\ldots,a_n \in S \right) \geq 1 - \frac{(2d+1)D}{\mathrm{card}(S)}$$

by the Schwartz-Zippel lemma 3.2. □

Before we prove the main theorem of this section we prove that the substitutions rarely allow that a gcd of higher degree occurs.

Lemma 3.5 ([10])**.** Let k be a field, $f_1,\ldots,f_r \in k[X_1,\ldots,X_n]$ polynomials with $\deg(f_i) \leq D$ for $1 \leq i \leq r$ and $\gcd(f_1,\ldots,f_r) = 1$. Furthermore, assume that $f_1(0,\ldots,0) \neq 0$. Then there exists a polynomial $\Delta \in k[Z_2,\ldots,Z_n]$ with $\deg(\Delta) \leq 2D^2$ such that for any elements $b_2,\ldots,b_n \in k$ with $\Delta(b_2,\ldots,b_n) \neq 0$ we have

$$\gcd_{1\leq i\leq r} (f_i(X_1, b_2X_1,\ldots,b_nX_1)) = 1 \,.$$

Proof. Since $f_1(0,\ldots,0) \neq 0$ it follows that X_1 doesn't divide $f_1(X_1, Z_2X_1,\ldots,Z_nX_1)$. Furthermore, we have $\gcd(f_1,\ldots,f_r) = 1$ by assumption. Thus it follows

$$\gcd_{1\leq i\leq r} (f_i(X_1, Z_2X_1,\ldots,Z_nX_1)) = 1$$

in $k[X_1, Z_2,\ldots,Z_n]$. Therefore there exists, by Bezout's identity, polynomials $s_1,\ldots,s_r \in k(Z_2,\ldots,Z_n)[X_1]$ with $\deg(s_i) < D$ such that

$$1 = \sum_{i=1}^{r} s_i f_i(X_1, Z_2X_1,\ldots,Z_nX_1) \,.$$

This yields a linear system over $k(Z_2,\ldots,Z_n)$ in $2D$ equations and rD unknowns. By Cramer's rule we can find a solution in $\frac{1}{\Delta(Z_2,\ldots,Z_n)}k[Z_2,\ldots,Z_n]$ where Δ is a determinant of a $m \times m$-Matrix, $m \leq 2D$, of coefficients of powers of X_1 in $f_i(X_1, Z_2X_1,\ldots,Z_nX_1)$. Hence $\deg(\Delta) \leq 2D^2$ and $b_2,\ldots,b_n \in k$ with $\Delta(b_2,\ldots,b_n) \neq 0$ implies

$$1 = \sum_{i=1}^{r} s_i(b_2,\ldots,b_n,X_1) f_i(X_1, b_2X_1,\ldots,b_nX_1)$$

and thus $\gcd_{1\leq i\leq r}(f_i(X_1, b_2X_1,\ldots,b_nX_1))$. □

Moreover, substitutions of the form $X_i \mapsto X_i + a_i$ have no influence on the irreducibility of polynomials.

Lemma 3.6. Let $f \in k[X_1,\ldots,X_n,Y]$ be a non-zero polynomial over a field k. For elements $a_1,\ldots,a_n \in k$ the polynomial

$$f(X_1 + a_1,\ldots,X_n + a_n, Y)$$

is irreducible if and only if f is irreducible.

Proof. The statement follows immediately from the fact that the map

$$k[X_1,\ldots,X_n,Y] \to k[X_1,\ldots,X_n,Y], \quad \begin{matrix} X_i & \mapsto X_i + a_i \\ Y & \mapsto Y \end{matrix} \tag{3.1}$$

is a k-algebra automorphism. □

In preparation of the proof of the effective Hilbert irreducibility theorem we state the following application of the Hensel lifting theorem.

Proposition 3.7. Let $g \in k[X_1, \ldots, X_n, Y]$ be an irreducible polynomial. Assume that $g(0, \ldots, 0, Y)$ is square free and $\mathrm{lc}_Y(g)(0, \ldots, 0) \neq 0$. Define for elements $b_2, \ldots, b_n \in k$

$$g_{\boldsymbol{b}}(X_1, Y) := g(X_1, b_2X_1, \ldots, b_nX_1, Y) \; .$$

Then there exists for each factor $h_{\boldsymbol{b}} \in k[[X_1]][Y]$ of $g_{\boldsymbol{b}}$ with $\mathrm{lc}_Y(h_{\boldsymbol{b}}) = \mathrm{lc}_Y(g_{\boldsymbol{b}})$ a factor $h \in k[[X_1, \ldots, X_n]][Y]$ of g with $\mathrm{lc}_Y(h) = \mathrm{lc}_Y(g)$ such that

$$h(X_1, b_2X_1, \ldots, b_nX_1, Y) = h_{\boldsymbol{b}}(X_1, Y) \; .$$

Proof. Remember that $k[[X_1, \ldots, X_n]]$ is a complete Noetherian local ring with maximal ideal $\mathfrak{m} = (X_1, \ldots, X_n)$. Let $h_{\boldsymbol{b}}(X_1, Y) \in k[[X_1]][Y]$ be a factor of $g_{\boldsymbol{b}}$ with $\mathrm{lc}_Y(h_{\boldsymbol{b}}) = \mathrm{lc}_Y(g_{\boldsymbol{b}})$. Notice that

$$g_{\boldsymbol{b}}(0, Y) = g(0, \ldots, 0, Y) \equiv g \bmod \mathfrak{m} \tag{3.2}$$

and that $\mathrm{lc}_Y(g)(0, \ldots, 0) \neq 0$ implies $\mathrm{lc}_Y(g) \notin \mathfrak{m}$. Furthermore, $h_{\boldsymbol{b}}(0, Y)$ is a non-zero factor of $g(0, \ldots, 0, Y)$. We write $g(0, \ldots, 0, Y) = h_{\boldsymbol{b}}(0, Y)\overline{h_{\boldsymbol{b}}}$ where $\overline{h_{\boldsymbol{b}}}$ is the corresponding cofactor. Since $g(0, \ldots, 0, Y)$ is square free $h_{\boldsymbol{b}}(0, Y)$ and $\overline{h_{\boldsymbol{b}}}$ are relatively prime. Therefore we can obtain by corollary 2.62 to the Hensel lifting a *unique* factor $h \in k[[X_1, \ldots, X_n]][Y]$ of g with $h(0, \ldots, 0, Y) = h_{\boldsymbol{b}}(0, Y)$ and $\mathrm{lc}_Y(h) = \mathrm{lc}_Y(g)$.

We claim that

$$h(X_1, b_2X_1, \ldots, b_nX_1, Y) = h_{\boldsymbol{b}}(X_1, Y) \; .$$

This can be seen as follows. Since $g_{\boldsymbol{b}}(0, Y) = g(0, \ldots, 0, Y)$ we can also consider the factorization $g_{\boldsymbol{b}}(0, Y) = h_{\boldsymbol{b}}(0, Y)\overline{h_{\boldsymbol{b}}}$ and apply on this factorization the Hensel lifting theorem. Since the lifted polynomials are unique we obtain as the lifted polynomial $h_{\boldsymbol{b}}$. Assume now that $h(X_1, b_2X_1, \ldots, b_nX_1, Y) \neq h_{\boldsymbol{b}}(X_1, Y)$. Then would $h(X_1, b_2X_1, \ldots, b_nX_1, Y)$ be another factor of $g_{\boldsymbol{b}}$ with

$$h(X_1, b_2X_1, \ldots, b_nX_1, Y) \equiv h_{\boldsymbol{b}}(0, Y) \bmod X_1$$

in contradiction to the uniqueness of the lifted polynomials! □

Now we can state the effective Hilbert irreducibility theorem under the additional constraint that for $f \in k[X_1, \ldots, X_n, Y]$ we have $\partial f/\partial Y \neq 0$. This is always satisfied if $\deg_Y(f) \geq 1$ and $\mathrm{char}(k) = 0$. The proof is based on the original paper by Kaltofen [10].

Theorem 3.8. Let k be a field, $f \in k[X_1, \ldots, X_n, Y]$ an irreducible polynomial with $\partial f/\partial Y \neq 0$ and δ the total degree of f. Pick random elements $a_1, \ldots, a_n, b_2, \ldots, b_n$ from a finite subset $S \subset k$. Then the probability

$$\mathrm{Prob}\left(f(a_1 + X_1, a_2 + b_2X_1, \ldots, a_n + b_nX_1, Y) \text{ irreducible in } k[X_1, Y] \right) \geq 1 - \frac{4\delta 2^\delta}{\mathrm{card}(S)} \; .$$

Proof. By lemma 3.4 the probability that $f(a_1,\dots,a_n,Y)$ is square free and $\mathrm{lc}_Y(f)(a_1,\dots,a_n) \neq 0$ is at least $1-(2d+1)D/\mathrm{card}(S)$ where $D := \deg_{X_1,\dots,X_n}(f)$ and $d := \deg_Y(f)$. Fix $a_1,\dots,a_n$ such that this is the case and set

$$g(X_1,\dots,X_n,Y) := f(X_1+a_1,\dots,X_n+a_n,Y)\ .$$

Notice that g is irreducible if and only if f is irreducible by lemma 3.6. For elements $b_2,\dots,b_n \in S$ set

$$\begin{aligned} g_{\boldsymbol{b}}(X_1,Y) &:= g(X_1,b_2X_1,\dots,b_nX_1,Y) \\ &= f(a_1+X_1,a_2+b_2X_1,\dots,a_n+b_nX_1,Y)\ . \end{aligned}$$

We have to prove that $g_{\boldsymbol{b}}$ is irreducible in $k[X_1,Y]$ with a probability of at least $1-4\delta 2^{\delta}/\mathrm{card}(S)$.

First we determine the probability that $g_{\boldsymbol{b}}$ remains irreducible in $k(X_1)[Y]$. By our assumption $g(0,\dots,0,Y) = f(a_1,\dots,a_n,Y)$ is square free and $\mathrm{lc}_Y(g)(0,\dots,0) = \mathrm{lc}_Y(f)(a_1,\dots,a_n) \neq 0$. Therefore by proposition 3.7 there exists for each factor $h_{\boldsymbol{b}} \in k[[X_1]][Y]$ of $g_{\boldsymbol{b}}$ with $\mathrm{lc}_Y(h_{\boldsymbol{b}}) = \mathrm{lc}_Y(g_{\boldsymbol{b}})$ a factor $h \in k[[X_1,\dots,X_n]][Y]$ of g with $\mathrm{lc}_Y(h) = \mathrm{lc}_Y(g)$ such that

$$h(X_1,b_2X_1,\dots,b_nX_1,Y) = h_{\boldsymbol{b}}(X_1,Y)\ .$$

Thus it is sufficient to show that for each factor $h \in k[[X_1,\dots,X_n]][Y]$ of g the polynomial $h_{\boldsymbol{b}}(X_1,Y) := h(X_1,b_2X_1,\dots,b_nX_1,Y)$ does not divide $g_{\boldsymbol{b}}$ in $k[X_1][Y]$. A sufficient condition is that $\deg_{X_1}(h_{\boldsymbol{b}}) > D = \deg_{X_1}(g_{\boldsymbol{b}})$. Then $g_{\boldsymbol{b}}$ remains irreducible in $k(X_1)[Y]$. We show that this is the case if a certain polynomial $\pi \in k[Z_2,\dots,Z_n]$ with degree at most $2D(2^{d-1}-1)$ does not vanish at $b_2,\dots,b_n \in S$. We fix the lead coefficients of the factors since $\mathrm{lc}_Y(g)$ and $\mathrm{lc}_Y(g_{\boldsymbol{b}})$ are units in $k[[X_1,\dots,X_n]]$ and $k[[X_1]]$ respectively and we are only interested in non-associated factors.

Let $h \in k[[X_1,\dots,X_n]][Y]$ be a factor of g with $\mathrm{lc}_Y(h) = \mathrm{lc}_Y(g)$ and let $\bar{h}$ be the corresponding cofactor such that $g = h\bar{h}$. We write with $\boldsymbol{X} := (X_1,\dots,X_n)$

$$h = \sum_{i=0}^{r}\sum_{\boldsymbol{\alpha}\in\mathbb{N}^n} h_{i,\boldsymbol{\alpha}}\boldsymbol{X}^{\boldsymbol{\alpha}}Y^i \quad \text{and} \quad \bar{h} = \sum_{i=0}^{s}\sum_{\boldsymbol{\alpha}\in\mathbb{N}^n} \bar{h}_{i,\boldsymbol{\alpha}}\boldsymbol{X}^{\boldsymbol{\alpha}}Y^i$$

where $s < d$ and $s+r = d$. We claim that there must exist a coefficient $h_{i,\boldsymbol{\alpha}}$ of h or a coefficient $\bar{h}_{i,\boldsymbol{\alpha}}$ of $\bar{h}$ for an index i with

$$D < |\boldsymbol{\alpha}| \leq 2D \quad \text{and} \quad \big(h_{i,\boldsymbol{\alpha}} \neq 0 \text{ or } \bar{h}_{i,\boldsymbol{\alpha}} \neq 0\big)\ .$$

Assume this were not the case. Then

$$g = \left(\sum_{i=0}^{r}\sum_{|\boldsymbol{\alpha}|\leq D} h_{i,\boldsymbol{\alpha}}\boldsymbol{X}^{\boldsymbol{\alpha}}Y^i\right)\left(\sum_{i=0}^{s}\sum_{|\boldsymbol{\alpha}|\leq D} \bar{h}_{i,\boldsymbol{\alpha}}\boldsymbol{X}^{\boldsymbol{\alpha}}Y^i\right) + \sum_{i=0}^{d}\sum_{\substack{|\boldsymbol{\beta}|\geq 2D+1 \\ \boldsymbol{\alpha}_1+\boldsymbol{\alpha}_2=\boldsymbol{\beta}}} h_{i,\boldsymbol{\alpha}_1}\bar{h}_{i,\boldsymbol{\alpha}_2}\boldsymbol{X}^{\boldsymbol{\beta}}Y^i\ .$$

Since $\deg_{\boldsymbol{X}}(g) \leq D$ the right sum vanishes and h and $\bar{h}$ can be considered as elements of $k[X_1,\dots,X_n][Y]$. Then g is reducible in $k[X_1,\dots,X_n][Y]$ in contradiction to f irreducible. Hence we can assume without loss of generality that there exists an $\boldsymbol{\alpha} \in \mathbb{N}^n$ and an integer i such that $h_{i,\boldsymbol{\alpha}} \neq 0$ with $D < |\boldsymbol{\alpha}| \leq 2D$ and $0 \leq i \leq r$. Then we can define

the polynomial

$$\delta_{i,\boldsymbol{\alpha}} := \sum_{|\boldsymbol{\beta}|=|\boldsymbol{\alpha}|} h_{i,\boldsymbol{\beta}} \boldsymbol{X}^{\beta} \neq 0$$

which is the coefficient of Y^i in h of degree $2D \geq |\boldsymbol{\alpha}| > D$ in $\boldsymbol{X}$.

If $b_2, \ldots, b_n \in S$ satisfy

$$\delta_{i,\boldsymbol{\alpha}}(X_1, b_2 X_1, \ldots, b_n X_1) \neq 0$$

we can guarantee that $h_{\boldsymbol{b}}$ has a non-zero coefficient of order $2D \geq |\boldsymbol{\alpha}| > D$ in X_1. Therefore $h_{\boldsymbol{b}}$ cannot be a polynomial dividing $g_{\boldsymbol{b}}$ in $k[X_1][Y]$. Thus the polynomial $\pi(Z_2, \ldots, Z_n)$ can be chosen as the product of the $\delta_{i,\boldsymbol{\alpha}}(1, Z_2, \ldots, Z_n) \neq 0$ over all possible factor candidates of h. Since g has at most d irreducible factors in $k[[X_1, \ldots, X_n]][Y]$ and we do not need to consider complementary factor combinations there are at most

$$\sum_{i=1}^{d-1} \binom{d-1}{i} = 2^{d-1} - 1$$

factors to refute. Hence $\deg(\pi) \leq 2D(2^{d-1} - 1)$ and we know that $\pi(b_2, \ldots, b_n) \neq 0$ guarantees that $g_{\boldsymbol{b}}$ has no factor $h_{\boldsymbol{b}}$ in $k[[X_1]][Y]$ with $\deg_{X_1}(h_{\boldsymbol{b}}) \leq \deg_{X_1}(g_{\boldsymbol{b}})$. Therefore $g_{\boldsymbol{b}}$ is irreducible in $k(X_1)[Y]$ if $\pi(b_2, \ldots, b_n) \neq 0$.

Finally we must refute a possible content in $k[X_1][Y]$. Let $l_i(X_1, \ldots, X_n)$ be the coefficient of Y^i in $g(X_1, \ldots, X_n, Y)$, $\deg(l_i) \leq D$. Then $l_d = \mathrm{lc}_Y(g)$ and thus $l_d(0, \ldots, 0) \neq 0$. Since f is irreducible we have $\gcd_{0 \leq i \leq d}(l_i) = 1$. By lemma 3.5 there exists a polynomial $\Delta \in k[Z_2, \ldots, Z_n]$ with $\deg(\Delta) \leq 2D^2$ such that for $b_2, \ldots, b_n \in S$ with $\Delta(b_2, \ldots, b_n) \neq 0$ it follows that $\gcd_{0 \leq i \leq d}(l_i(X_1, b_2 X_1, \ldots, b_n X_1)) = 1$. For such $b_2, \ldots, b_n \in S$ it follows that $g_{\boldsymbol{b}}$ cannot have a non-trivial content with respect to Y, i.e., a factor in $k[X_1]$.

We conclude that we have to avoid zeros of $\pi\Delta$. For randomly chosen $b_2, \cdots, b_n \in S$ we have $\pi\Delta(b_2, \cdots, b_n) \neq 0$ with a probability of at least $1 - (\deg(\pi) + \deg(\Delta))/\mathrm{card}(S)$ by the Schwartz-Zippel lemma 3.2. Together with the probability that $f(a_1, \ldots, a_n, Y)$ is square free and $\mathrm{lc}_Y(f)(a_1, \ldots, a_n) \neq 0$ it follows that $g_{\boldsymbol{b}}$ is irreducible with a probability of at least

$$\left(1 - \frac{(2d+1)D}{\mathrm{card}(S)}\right)\left(1 - \frac{2D(2^{D-1} - 1) + 2D^2}{\mathrm{card}(S)}\right) \geq 1 - \frac{4\delta 2^{\delta} - 3d}{\mathrm{card}(S)} \geq 1 - \frac{4\delta 2^{\delta}}{\mathrm{card}(S)}$$

where $\delta := dD = \deg(f)$. □

For every non-zero polynomial f the condition $\partial f / \partial Y \neq 0$ is satisfied if k is a field of characteristic 0. For characteristic $p > 0$ one can prove that without the assumption about the derivative the theorem is still correct if k is a perfect field, in particular if every element of k is a pth power. But our black box factorization algorithm succeeds only with a controllably high probability if k is an field of characteristic zero. Therefore we only state the general theorem and skip the proof.

Theorem 3.9 (Effective Hilbert irreducibility theorem)**.** Let k be a perfect field, $f \in k[X_1, \ldots, X_n, Y]$ an irreducible polynomial with degree δ. Pick random elements $a_1, \ldots, a_n, b_2, \ldots, b_n$ from a finite subset $S \subset k$. Then

$$\text{Prob}\ \big(\ f(a_1 + X_1, a_2 + b_2X_1, \ldots, a_n + b_nX_1, Y) \text{ irreducible in } k[X_1, Y]\ \big) \geq 1 - \frac{4\delta 2^\delta}{\text{card}(S)}\ .$$

Proof. If $\text{char}(k) = 0$ this is theorem 3.8. If $\text{char}(k) > 0$ see [10]. □

3.2 Factor degree pattern

In the previous section we determined the probability that for randomly chosen elements $a_1, \ldots, a_n, b_2, \ldots, b_n$ the image $f(a_1 + X_1, a_2 + b_2X_1, \ldots, a_n + b_nX_1, Y)$ of an irreducible polynomial $f \in k[X_1, \ldots, X_n, Y]$ remains irreducible. The question that now arises is what happens with the image $f(a_1 + X_1, a_2 + b_2X_1, \ldots, a_n + b_nX_1, Y)$ of a *reducible* polynomial?

Consider the factorization $f = g_1^{e_1} \cdots g_r^{e_r}$ of f in pairwise non-associated irreducible factors g_i with $d_i = \deg(g_i) \geq 1$ and $e_i \geq 1$. We call the lexicographically ordered n-tuple $((d_{i_1}, e_{i_1}), \ldots, (d_{i_r}, e_{i_r}))$ the *factor degree pattern* of f.

The images $g_i(a_1 + X_1, a_2 + b_2X_1, \ldots, a_n + b_nX_1, Y)$ remain irreducible if they depend on Y by the effective Hilbert irreducibility theorem 3.9. But they have not necessarily the same degree and they can become associated. Thus the interesting question is whether the factor degree pattern of the image $f(a_1 + b_1X_1, \ldots, a_n + b_nX_1, Y)$ coincides with the factor degree pattern of $f(X_1, \ldots, X_n, Y)$. Since we can only apply the effective Hilbert irreducibility theorem on those factors that depend on Y we need the following notation. The *primitive part* of a polynomial f with respect to some indeterminate Y is the polynomial divided by the gcd of all coefficients with respect to Y (the *content*) and denoted by $\text{pp}_Y(f)$. If the content of a f is an unit then f is called *primitive*.

Theorem 3.10 (Factor degree pattern [12])**.** Let $f \in k[X_1, \ldots, X_n, Y]$ be a polynomial over a perfect field k. Denote by δ the total degree of f and pick random elements $a_1, \ldots, a_n, b_2, \ldots, b_n$ from a finite subset $S \subset k$ and set

$$f_2 := f(a_1 + X_1, a_2 + b_2X_1, \ldots, a_n + b_nX_1, Y)\ .$$

Then

$$\text{Prob}\ \big(\ \text{pp}_Y(f) \text{ and } \text{pp}_Y(f_2) \text{ have the same factor degree pattern}\ \big) \geq 1 - \frac{4\delta 2^\delta + \delta^3}{\text{card}(S)}\ .$$

Proof. Let $f = \prod_{i=1}^r g_i^{e_i}$ be a factorization of f in pairwise non-associated irreducible factors g_i with $\delta_i = \deg(g_i)$ and $e_i \geq 1$ for $i = 1, \ldots, r$. First consider the factors g_i with $\deg_Y(g_i) > 0$. By the effective Hilbert irreducibility theorem 3.9

$$g_{i,2} := g_i(a_1 + X_1, a_2 + b_2X_1, \ldots, a_n + b_nX_1, Y)$$

remains irreducible in $k[X_1, Y]$ with a probability of at least $1 - 4\delta_i 2^{\delta_i}/\mathrm{card}(S)$. It remains to determine the probability that $\deg(g_{i,2}) = \delta_i$ and that the $g_{i,2}$ are pairwise non-associated.

We start with the degree. Let $\boldsymbol{A} := (A_1, \ldots, A_n)$ and $\boldsymbol{B} := (B_2, \ldots, B_n)$ be two families of indeterminates and define

$$h_i(X_1, Y, \boldsymbol{A}, \boldsymbol{B}) := g_i(A_1 + X_1, A_2 + B_2 X_1, \ldots, A_n + B_n X_1, Y) \; .$$

Notice that $h_i(X_1, Y, a_1, \ldots, a_n, b_2, \ldots, b_n) = g_{i,2}$. Clearly $\deg_{X_1,Y}(h_i) = \delta_i$ for $i = 1, \ldots, r$ and thus there exists in each h_i a non-zero coefficient $\pi_i(B_2, \ldots, B_n)$ of a monomial $X_1^j Y^k$ with $j + k = \delta_i$. Then $\deg(\pi_i) \leq \delta_i$ and $\pi_i(b_2, \ldots, b_n) \neq 0$ implies $\deg(g_{i,2}) = \delta_i$. The probability that $\pi_i(b_2, \ldots, b_n) \neq 0$ is at least

$$1 - \deg(\pi_i)/\mathrm{card}(S) \geq 1 - \delta_i/\mathrm{card}(S)$$

by the Schwartz-Zippel lemma 3.2.

Now we have to determine the probability that the g_i are pairwise non-associated. We claim that the h_i are pairwise non-associated in $k(\boldsymbol{A}, \boldsymbol{B})[X_1, Y]$. Assume this were not the case. Then there exists integers i and j, $i \neq j$, and polynomials $s_i, s_j \in k[\boldsymbol{A}, \boldsymbol{B}]$ with $\gcd(s_i, s_j) = 1$ such that $s_i h_i = s_j h_j$. Since $\deg s_i \geq 1$ it follows that s_i divides h_j and thus h_j is reducible in $k[X_1, Y, \boldsymbol{A}, \boldsymbol{B}]$. We write $h_j = h_j^{(1)} h_j^{(2)}$. But then would

$$\begin{aligned} g_j(X_1, \ldots, X_n, Y) &= h_j(X_1, Y, 0, X_2 - B_2 X_1, \ldots, X_n - B_n X_1, B_2, \ldots, B_n) \\ &= (h_j^{(1)} h_j^{(2)})(X_1, Y, 0, X_2 - B_2 X_1, \ldots, X_n - B_n X_1, B_2, \ldots, B_n) \end{aligned}$$

be a non-trivial factorization of g_j in $k[X_1 \ldots, X_n, Y]$ in contradiction to g_j irreducible. Therefore the h_k are pairwise non-associated in $k(\boldsymbol{A}, \boldsymbol{B})[X_1, Y]$.

Hence there exists in h_i and h_j, $i \neq j$, coefficients $h_i^{(\alpha_1,\alpha_2)}$ and $h_j^{(\alpha_1,\alpha_2)}$ of $X_1^{\alpha_1} Y^{\alpha_2}$ such that

$$h_j^{(\alpha_1,\alpha_2)} h_i - h_i^{(\alpha_1,\alpha_2)} h_j \neq 0$$

(otherwise h_i and h_j would be associated). In particular there exists two additional coefficients $h_i^{(\beta_1,\beta_2)}$ and $h_j^{(\beta_1,\beta_2)}$ of $X_1^{\beta_1} Y^{\beta_2}$ in h_i and h_j such that

$$\tau_{i,j} := h_j^{(\alpha_1,\alpha_2)} h_i^{(\beta_1,\beta_2)} - h_i^{(\alpha_1,\alpha_2)} h_j^{(\beta_1,\beta_2)} \neq 0 \; .$$

$\tau_{i,j}$ is a polynomial in $k[\boldsymbol{A}, \boldsymbol{B}]$ and $\tau_{i,j}(a_1, \ldots, a_n, b_2, \ldots, b_n) \neq 0$ implies that $g_{i,2}$ and $g_{j,2}$ are not associated. Since $\deg(\tau_{i,j}) \leq \delta_i + \delta_j$ the probability that $g_{i,2}$ and $g_{j,2}$ are not associated is at least $1 - (\delta_i + \delta_j)/\mathrm{card}(S)$ by the Schwartz-Zippel lemma 3.2.

Finally we consider the case $\deg_Y(g_i) = 0$. Then $g_{i,2}$ is a divisor of the content of f_2 with respect to Y. Thus it is sufficient that $g_{i,2}$ is not identical zero. A condition for this is that the total degree of $g_{i,2}$ gets preserved. By the same arguments as in the first case this happens with a probability of at least $1 - \delta_i/\mathrm{card}(S)$.

In summary it follows that that the factor degree pattern is preserved with a probability not less than

$$\begin{aligned}
&1 - \left(\underbrace{\sum_{i=1}^{r} \frac{4\delta_i 2^{\delta_i}}{\mathrm{card}(S)}}_{\text{irreducible}} + \underbrace{\sum_{i=1}^{r} \frac{\delta_i}{\mathrm{card}(S)}}_{\text{degree preserved}} + \underbrace{\sum_{1 \le i < j \le r} \frac{\delta_i + \delta_j}{\mathrm{card}(S)}}_{\text{non-associated}} \right) \\
\geq & 1 - \frac{1}{\mathrm{card}(S)} \left(4\delta 2^{\delta} + \delta + \frac{\delta(\delta-1)}{2}\delta \right) \\
\geq & 1 - \frac{4\delta 2^{\delta} + \delta^3}{\mathrm{card}(S)}
\end{aligned}$$

□

With this theorem we can probabilistically guarantee that the factor degree patterns of f and $f(a_1 + X_1, a_2 + b_2 X_1, \ldots, a_n + b_n X_1, Y)$ coincide if f is primitive with respect to Y. But what happens if f is not primitive? The idea is that we modify f in such a way that the image is primitive with a high probability and has the same factor degree pattern as f. The following lemmas show that we can in fact define such an image of f.

Lemma 3.11. Let $f \in k[X_1, \ldots, X_n, Y]$ be a polynomial over a field k with total degree δ and pick random elements $c_1, \ldots, c_n$ from a finite subset $S \subset k$. Then

$$\mathrm{Prob}\left(\mathrm{lc}_Y(f(X_1 + c_1 Y, \ldots, X_n + c_n Y, Y)) \in k \right) \geq 1 - \frac{\delta}{\mathrm{card}(S)} .$$

Proof. For indeterminates $C_1, \ldots, C_n$ we define

$$\bar{f}(C_1, \ldots, C_n, X_1, \ldots, X_n, Y) := f(X_1 + C_1 Y, \ldots, X_n + C_n Y, Y)$$

and $\pi := \mathrm{lc}_Y(\bar{f}) \in k[C_1, \ldots, C_n]$. Then $\deg(\pi) \leq \delta$ and if $\pi(c_1, \ldots, c_n) \neq 0$ for $c_1, \ldots, c_n \in S$ it follows $\mathrm{lc}_Y(f(X_1 + c_1 Y, \ldots, X_n + c_n Y, Y)) \in k$ and by the Schwartz-Zippel lemma 3.2 it follows the statement. □

Moreover, substitutions of the form $X_i \mapsto X_i + b_i Y + a_i$ have no influence on the factor degree pattern of f.

Lemma 3.12. Let $f \in k[X_1, \ldots, X_n, Y]$ be a non-zero polynomial over a field k. For elements $a_i, b_i \in k$, $i = 1, \ldots, n$

$$f(X_1 + b_1 Y + a_1, \ldots, X_n + b_n Y + a_n, Y)$$

and f have the same factor degree patterns.

Proof. The statement follows immediately from the fact that the map

$$k[X_1, \ldots, X_n, Y] \to k[X_1, \ldots, X_n, Y], \quad \begin{matrix} X_i & \mapsto X_i + b_i Y + a_i \\ Y & \mapsto Y \end{matrix} \tag{3.3}$$

is a k-algebra automorphism. □

Therefore we can construct for a multivariate polynomial f a bivariate polynomial f_2 such that the factor degree patterns of f and f_2 coincide with a controllable high probability.

As a final result we state the substitution which is used in our factorization algorithm and the probability that the factor degree patterns coincide for this substitution.

Corollary 3.13. Let $f \in k[X_1, \ldots, X_n, Y]$ be a non-zero polynomial over a perfect field k. Denote by δ the total degree of f and pick randomly chosen elements $a_1, \ldots, a_n$, $b_2, \ldots, b_n, c_1, \ldots, c_n$ and a_Y, b_Y from a finite subset $S \subset k$ and set

$$f_2 = f(a_1 + X_1 + c_1 Y, a_2 + b_2 X_1 + c_2 Y, \ldots, a_n + b_n X_1 + c_n Y, a_Y + b_Y X_1 + Y)$$

Then

$$\text{Prob}\left(f \text{ and } f_2 \text{ have the same factor degree pattern} \right) \geq 1 - \frac{\delta + 4\delta 2^\delta + \delta^3}{\text{card}(S)} . \quad (3.4)$$

Furthermore, for the factorization $\prod_{i=1}^r g_{2,i}^{e_i}$ of f_2 with $g_{2,i} \in k[X_1, Y]$ we have

$$\text{Prob}\left(\begin{array}{l} \deg(g_{2,i}(X_1, 0)) = \deg(g_{2,i}) \text{ for } i = 1, \ldots, r \text{ and} \\ \gcd(g_{2,i}(X_1, 0), g_{2,j}(X_1, 0)) = 1 \text{ for } 1 \leq i < j \leq r \end{array} \right) \geq 1 - \frac{\delta^2}{\text{card}(S)} . \quad (3.5)$$

Proof. First set $\tilde{f}_1(X_1, \ldots, X_n, Y) := f(X_1 + c_1 Y, \ldots, X_n + c_n Y, Y)$. By lemma 3.11 we have $\text{lc}_Y(\tilde{f}_1) \in k$ with a probability of at least $1 - \delta/\text{card}(S)$ and therefore $\tilde{f}_1$ is primitive in Y with the same probability. Assume now that is true. By lemma 3.12 the factor degree patterns of f and $\tilde{f}_1$ coincide and if we set

$$\begin{aligned} \tilde{f}_2(X_1, Y) &:= \tilde{f}_1(a_1 + X_1, a_2 + b_2 X_1, \ldots, a_n + b_n X_1, Y) \\ &= f(a_1 + X_1 + c_1 Y, a_2 + b_2 X_1 + c_2 Y, \ldots, a_n + b_n X_1 + c_n Y, Y) \end{aligned}$$

it follows that the factor degree patterns of $\tilde{f}_1$ and $\tilde{f}_2$ coincide with a probability of at least $1 - (4\delta 2^\delta + \delta^3)/\text{card}(S)$ by theorem 3.10. Hence

$$\text{Prob}\left(f \text{ and } \tilde{f}_2 \text{ have the same factor degree pattern} \right) \geq 1 - \frac{\delta + 4\delta 2^\delta + \delta^3}{\text{card}(S)} .$$

Again assume that this is true and write $\tilde{f}_2 = \prod_{i=1}^r \tilde{g}_{2,i}^{e_i}$ for the factorization of $\tilde{f}_2$. Then

$$\tilde{g}_{2,i} = g_i(a_1 + X_1 + c_1 Y, a_2 + b_2 X_1 + c_2 Y, \ldots, a_n + b_n X_1 + c_n Y, Y)$$

where g_i is the corresponding factor of f. For $i = 1, \ldots, r$ define

$$h_{2,i}(X_1, Y, A_Y, B_Y) := g_i(a_1 + X_1 + c_1 Y, a_2 + b_2 X_1 + c_2 Y, \ldots, a_n + b_n X_1 + c_n Y, A_Y + B_Y X_1 + Y) .$$

For $g_{2,i} := h_{2,i}(X_1, Y, a_Y, b_Y) \in k[X_1, Y]$ it follows easily that the $g_{2,i}$ are irreducible and pairwise not associated. Therefore $\prod_{i=1}^r g_{2,i}^{e_i}$ is the factorization of f_2 and the factor degree patterns of $\tilde{f}_2$ and f_2 coincide. This shows statement (3.4).

Now we want to show statement (3.5). By our previous construction we have $g_{2,i}(X_1, 0) = h_{2,i}(X_1, 0, a_Y, b_Y)$ for $1 \leq i \leq r$. Set $\pi_i(B_Y) := \text{lc}_{X_1}(h_{2,i})$ for $i = 1, \ldots, r$ and

$$\sigma_{i,j}(A_Y, B_Y) := \text{res}_{X_1}(h_{2,i}(X_1, 0, A_Y, B_Y), h_{2,j}(X_1, 0, A_Y, B_Y))$$

for $1 \leq i < j \leq r$. Then $\pi_i(b_Y) \neq 0$ implies $\deg(g_{1,i}) = \deg(g_{2,i})$ and $\sigma_{i,j}(a_Y, b_Y) \neq 0$ implies $\operatorname{res}_{X_1}(g_{2,i}(X_1, 0), g_{2,j}(X_1, 0)) \neq 0$ and thus $\gcd(g_{2,i}(X_1, 0), g_{2,j}(X_1, 0)) = 1$ by corollary 2.51. Therefore it is sufficient to determine the probability that

$$\prod_{i=1}^{r} \pi_i(b_Y) \prod_{1 \leq i < j \leq r} \sigma_{i,j}(a_Y, b_Y) = 0 \; .$$

Set $\delta_i := \deg(g_{2,i})$ for $i = 1, \ldots, r$. Now $\deg(\pi_i) \leq \delta_i$ for $i = 1, \ldots, r$ and $\deg(\sigma_{i,j}) \leq 2\delta_i\delta_j$ for $1 \leq i < j \leq r$. Therefore

$$\deg\Big(\prod_{i=1}^{r} \pi_i \prod_{1 \leq i < j \leq r} \sigma_{i,j}\Big) \leq \sum_{i=1}^{r} \delta_i + \sum_{1 \leq i < j \leq r} 2\delta_i\delta_j \leq (\delta_1 + \cdots + \delta_r)^2 \leq \delta^2 \; .$$

Hence it follows statement (3.5) by the Schwartz-Zippel lemma 3.2. □

Remark 3.14. The proof of statement (3.5) is based on [12], proof of theorem 6.1.

Chapter 4

Black box factorization

Finally we describe the black box factorization algorithm proposed by Kaltofen and Trager [13].

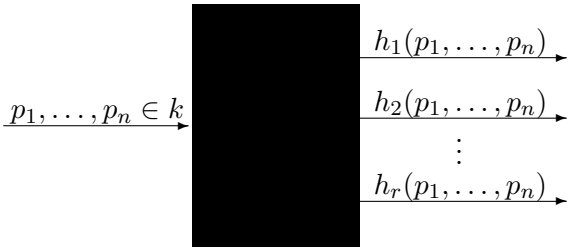
Algorithm 4.1 Black box polynomial factorization

Input:

A non-zero polynomial $f \in k[X_1, \ldots, X_n]$ given by a black box $\mathcal{B}_f$, where k is a field of characteristic 0, and the total degree d of f. We also assume that we have an efficient polynomial time factorization algorithm for $k[X_1, X_2]$. Furthermore a failure probability $\epsilon \ll 1$ is part of the input.

Output:

Assume $f = \prod_{i=1}^{r} h_i^{e_i}$ is the factorization of f in irreducible, pairwise non-associated polynomials $h_i \in k[X_1, \ldots, X_n]$ with multiplicity $e_i \geq 1$. First we return positive integers $\tilde{e}_1, \ldots, \tilde{e}_{\tilde{r}}$ such that $\tilde{e}_i = e_i$ for $i = 1, \ldots, r$ and $\tilde{r} = r$ with a probability of at least $1 - \epsilon$. Second we return the following output program. The program accepts as input n arbitrary elements $p_1, \ldots, p_n \in k$ and returns the values $h_1(p_1, \ldots, p_n), \ldots, h_r(p_1, \ldots, p_n) \in k$:



Notice that the h_i are determined only up to a multiple in k. The constructed program once and for all chooses an associate for each factor h_i and, for repeated invocations with different arguments, returns the value of that associate. Notice also that the failure probability applies to the construction and not to the execution of

the program. That is, with probability of at least $1-\epsilon$ the output program is correct; a correct program always produces the true values of the factors.

Step 1:

Pick randomly chosen elements $a_1, \ldots, a_n, b_2, \ldots, b_n, c_1, c_3, \ldots, c_n$ from a sufficiently large finite subset $S \subset k$ and compute by standard interpolation the following interpolation polynomial:

$$\begin{aligned} f_2(X_1, X_2) := \quad & f(X_1 + c_1X_2 + a_1, b_2X_1 + X_2 + a_2, b_3X_1 + c_3X_2 + a_3, \\ & \ldots, b_nX_1 + c_nX_2 + a_n) \end{aligned}$$

Step 2:

Factor $f_2(X_1, X_2)$ in $k[X_1, X_2]$ such that

$$f_2(X_1, X_2) = \prod_{i=1}^{\tilde{r}} g_{2,i}(X_1, X_2)^{\tilde{e}_i} \ .$$

With a probability not less than $1 - \epsilon$ we have $r = \tilde{r}$ and $e_i = \tilde{e}_i$ for all $1 \leq i \leq r$. Assume that this is all true. Otherwise an incorrect output program will be produced.

Step 3:

Assign

$$g_{1,i}(X_1) := g_{2,i}(X_1, 0) \quad , \quad 1 \leq i \leq r \ .$$

Check whether $\gcd(g_{1,i}, g_{1,j}) = 1$, $1 \leq i < j \leq r$, and $\deg(g_{2,i}) = \deg(g_{1,i})$, $1 \leq i \leq r$. If one check fails, return "failure". Then we have found that our chosen elements in step 1 were unlucky.

Now set

$$f_1(X_1) := f_2(X_1, 0) = f(X_1 + a_1, b_2X_1 + a_2, \ldots, b_nX_1 + a_n) = \prod_{i=1}^{r} g_{1,i}(X_1)^{e_i} \ .$$

We use the $g_{1,i}$ to uniquely enumerate the factors of f. Our associated choices (see output specifications) then satisfy

$$h_i(X_1 + a_1, b_2X_1 + a_2, \ldots, b_nX_1 + a_n) = g_{1,i}(X_1) \ .$$

Step 4:

This step constructs the output program for the evaluation of the h_i at $p_1, \ldots, p_n$ as described in the output specifications. First, the information computed so far is "hardwired" into that program. Then the following steps 4.1, 4.2, and 4.3 are appended to the program.

Step 4.1:

By standard interpolation compute

$$\begin{aligned}\bar{f}(X_1, Y) := \quad & f(X_1 + a_1, Y(p_2 - b_2(p_1 - a_1) - a_2) + b_2 X_1 + a_2, \\ & \ldots, Y(p_n - b_n(p_1 - a_1) - a_n) + b_n X_1 + a_n)\end{aligned} .$$

Notice that $\bar{f}(p_1 - a_1, 1) = f(p_1, \ldots, p_n)$ and $\bar{f}(X_1, 0) = f_1(X_1)$.

Step 4.2:

By Hensel lifting we obtain a factorization

$$\bar{f}(X_1, Y) \equiv \prod_{i=1}^{r} \bar{g}_i(X_1, Y)^{e_i} \bmod Y^{d+1} \quad \text{with } \bar{g}_i(X_1, 0) = g_{1,i}(X_1) \tag{4.1}$$

For all $1 \leq i \leq r$ test whether $\bar{g}_i$ divides $\bar{f}$. If at least one test fails return "failure". We have then discovered that the factor degree pattern of f and f_2 disagree.

Step 4.3:

For $i := 1, \ldots, r$ do:
return $g_i(p_1 - a_1, 1)$ as $h_i(p_1, \ldots, p_n)$

Step 5:

return $(\tilde{e}_1, \ldots, \tilde{e}_{\tilde{r}})$ and the program constructed in step 4.

Remark 4.2. The Black box polynomial factorization algorithm is a Monte Carlo algorithm.

First we prove the correctness, analyze the failure probability of the algorithm and then we prove the running time of the algorithm.

Theorem 4.3 (Correctness and failure probability)**.** The black box polynomial factorization algorithm 4.1 works correctly and if the cardinality of the set S in step 1 is chosen

$$\operatorname{card}(S) \geq 6 \deg(f) 2^{\deg(f)} / \epsilon,$$

then the algorithm succeeds with probability not less than $1 - \epsilon$ and the resulting program always correctly evaluates all irreducible factors of f.

Proof. Denote by δ the total degree of f. By corollary 3.13 the factor degree patterns of f and f_2 coincide in step 2 with a probability of at least $1 - (\delta + 4\delta 2^{\delta} + \delta^3)/\operatorname{card}(S)$. By the same corollary it follows that the probability that one of the checks in step 3 returns "failure" is less than $\delta^2/\operatorname{card}(S)$. Assume all this is true. Notice that then $\deg(g_{2,i}) = \deg(g_{1,i})$ for $i = 1, \ldots, r$ and this implies $\operatorname{lc}_{X_1}(f_2) \in k$ and thus also $\operatorname{lc}_{X_1}(\bar{f}) \in k$. Since $k[Y]$ is a Noetherian domain with maximal ideal (Y) we can apply our Hensel lifting algorithm 2.64 to the factorization

$$\bar{f}(X_1, Y) \equiv f_1(X_1) \equiv \prod_{i=1}^{r} g_{1,i}(X_1)^{e_i} \bmod Y$$

such that we obtain the factorization (4.1). If we now choose card(S) such that

$$\frac{\delta + 4\delta 2^{\delta} + \delta^3 + \delta^2}{\epsilon} = \frac{4\delta 2^{\delta} + \delta(1+\delta) + \delta(\delta^2)}{\epsilon} \leq \frac{6\delta 2^{\delta}}{\epsilon} \leq \mathrm{card}(S)$$

we can guarantee that the output program is correct (and step B returns never a failure) with a probability of at least $1 - \epsilon$. □

Theorem 4.4. The black box polynomial factorization algorithm 4.1 can construct its output program in polynomially many arithmetic steps as a function of n and $\deg(f)$ and an additional single polynomial factorization in $k[X_1, X_2]$. It requires $O(\deg(f)^2)$ calls to the black box for f. The output program can be executed in $\mathcal{O}((\log(\deg(f))+1)\deg(f)^3)$ arithmetic steps and $\mathcal{O}(\deg(f)^2)$ calls to the black box for f.

Proof. Each bivariate interpolation of f_2 and $\bar{f}$ requires $\mathcal{O}(\deg(f)^2)$ black box evaluations. Then the algorithm needs to factor f_2 which can be accomplished by our assumption in polynomial time in the size of $\deg(f)$ and the size of the coefficients of f. The dominating additional work of the output program is step 4.2, which can be accomplished in $\mathcal{O}((\log(\deg(f))+1)\deg(f)^3)$ arithmetic operations by theorem 2.65. □

Chapter 5

Closing remarks

In Mathematics it often occurs that after a novel idea gets introduced one notices that it is in fact just a rediscovery. A very specific case of this is Hensel lifting. It was introduced in computer algebra by Zassenhaus in 1969, who referred to the in 1908 published Hensel's Lemma. In this thesis it has been shown that the Hensel lifting is in fact a special case of Newton Iteration, an over 300 years old technique. But we have to note that the necessary algebraic concepts were first introduced in the 19th and 20th century. I hope that the link between Hensel lifting and Newton iteration was for the reader as fascinating as it was for me.

As for the black box factorization algorithm I believe it is not only a theoretically efficient solution for the problem of factoring multivariate polynomials but also very suitable for use in practice. For example the output program can be easily distributed, due to the small space requirement of a black box, to a network of asynchronous parallel processors and therefore evaluated in parallel. This enables for instance the computation of the sparse representation of the factors in parallel. Moreover, I want to remark that the algorithm can also be adapted for finite coefficient fields with sufficiently high characteristic. We just have to take into account that the failure probability is bounded below by the characteristic of the field.

Also with the black box approach Kaltofen and Trager were able to solve the gcd and the numerator / denominator problem for multivariate polynomials in random polynomial time [13]. This suggests that the black box approach could prove beneficial for other computational problems as well.

Bibliography

[1] M. Ben-Or and P. Tiwari. A deterministic algorithm for sparse multivariate polynomial interpolation. *Proc. 20th Annual ACM Symp. Theory Comp.*, pages 301–309, 1988.

[2] E. R. Berlekamp. Factoring polynomials over finite fields. *Bell Systems Tech. J.*, 46:1853–1859, 1967. Republished in revised form in: E. R. Berlekamp, Algebraic Coding Theory, Chapter 6, McGraw-Hill Publ., New York 1968.

[3] E. R. Berlekamp. Factoring polynomials over large finite fields. *Math. Comp.*, 24:713–735, 1970.

[4] N. Bourbaki. *Algebra II Chapters 4-7*, chapter IV, pages 31–32. Elements of mathematics. Springer-Verlag, 1989.

[5] N. Bourbaki. *Commutative Algebra Chapters 1-7*, page 392. Elements of mathematics. Springer-Verlag, 1989.

[6] N. Bourbaki. *Commutative Algebra Chapters 1-7*, pages 200–205 and 392. Elements of mathematics. Springer-Verlag, 1989.

[7] D. Eisenbud. *Commutative Algebra: With a View Toward Algebraic Geometry*, pages 181–182. Number 150 in Graduate Texts in Mathematics. Springer-Verlag, 1995.

[8] K. Hensel. *Theorie der algebraischen Zahlen*, chapter 4. Teubner, Leipzig, 1908.

[9] D. Hilbert. Über die Irreduzibilität ganzer rationaler Funktionen mit ganzahligen Koeffizienten. *J. reine angew. Math.*, 110:104–129, 1892.

[10] E. Kaltofen. Effective Hilbert Irreducibility. *Information and Control*, 66:123–137, 1985.

[11] E. Kaltofen. Polynomial-time reductions from multivariate to bi- and univariate integral polynomial factorization. *SIComp*, 14(2):469–489, 1985.

[12] E. Kaltofen. Factorization of polynomials given by straight-line programs. In S. Micali, editor, *Randomness and Computation*, volume 5 of *Advances in Computing Research*, pages 375–412. JAI Press Inc., 1989.

[13] E. Kaltofen and B. Trager. Computing with polynomials given by black boxes for their evaluations: Greatest common divisors, factorization, separation of numerators and denominators. *J. Symbolic Comput.*, 9(3):301–320, 1990.

[14] L. Kronecker. Grundzüge einer arithmetischen Theorie der algebraischen Grössen. *J. reine angew. Math*, 92:1–122, 1882.

[15] S. Landau. Factoring polynomials over algebraic number fields. *SIAM J. Comp.*, 14:184–195, 1985.

[16] A. K. Lenstra and H. W. Lenstra. Factoring polynomials with rational coefficients. *Math. Ann.*, pages 515–534, 1982.

[17] I. Newton. *Arithmetica Universalis, 2nd ed.* 1728. Reprinted in The Mathematical Works of Isaac Newton, vol. 2, D. T. Whiteside, ed., Johnson Reprint Corp., New York, 1967.

[18] B. L. van der Waerden. *Modern Algebra.* F. Ungar Publ. Co., New York, 1953.

[19] J. von zur Gathen and J. Gerhard. *Modern Computer Algebra - Third Edition*, chapter 6. Cambridge University Press, 2013.

[20] J. von zur Gathen and J. Gerhard. *Modern Computer Algebra - Third Edition*, chapter 9. Cambridge University Press, 2013.

[21] H. Zassenhaus. On Hensel factorization I. *J. Number Theory*, 1:291–311, 1969.