

Paul Breiding, Kathlén Kohn and Bernd Sturmfels

# Metric Algebraic Geometry

Springer Nature



# Preface

This book grew out of the lecture notes we developed for the Oberwolfach seminar on metric algebraic geometry which was held in the week of May 29 to June 2, 2023. Each of the three lecturers presented five of the 15 chapters. The lectures were supplemented by intense working sessions and inspiring evening discussions.

In the early 19th century, there was no difference between algebraic and differential geometry. The two were part of the same subject. Geometers studied natural properties of curves and surfaces in 3-space, such as curvature, singularities, and defining equations. In the 20th century, the threads diverged. The standard mathematics curriculum now offers algebraic geometry and differential geometry in rather disconnected courses.

In the present days, the geometry of data requires us to rethink that schism. Many applied problems center around metric questions, such as optimization with respect to distances. These require tools from different areas in geometry, algebra and analysis.

Metric algebraic geometry offers a path towards integration. This term is a neologism which joins the names metric geometry and algebraic geometry. It first appeared in the title of Madeleine Weinstein's PhD dissertation (UC Berkeley 2021). Building on classical foundations, the field embarks towards a new paradigm that combines concepts from algebraic geometry and differential geometry, with the goal of developing practical tools for the 21st century.

Many problems in the sciences lead to solving polynomial equations over the real numbers. The solution sets are real algebraic varieties. Understanding distances, volumes and angles – in short, understanding metric properties – of varieties is important for modeling and analyzing data. Other metric problems in optimization and statistics involve, for instance, minimizing the Euclidean distance from a variety to a given data point. Furthermore, in topological data analysis, computing the homology of a submanifold depends on curvature and on bottlenecks. Related metric questions arise in machine learning, in the geometry of computer vision, in learning varieties from data, and in the study of Voronoi cells.

This book addresses a wide audience of researchers and students, who will find it useful for seminars or self-study. It can serve as the text for a one-semester course at the graduate level. The key prerequisite is a solid foundation in undergraduate mathematics, especially in algebra, geometry and numerics. Course work in statistics and computer science, as well as experience with mathematical software, are helpful.

We hope that you will enjoy this book, and we invite you to develop your own perspective, and to share your feedback.

Oberwolfach,  
June 2023

*Paul Breiding  
Kathlén Kohn  
Bernd Sturmfels*



## Acknowledgements

The authors of this book wish to wholeheartedly thank the following 21 enthusiastic and hard-working participants of the Oberwolfach workshop on Metric Algebraic Geometry:

Patience Ablett  
Yueqi Cao  
Nick Dewaele  
Mirte van der Eyden  
Luca Fiorindo  
Sofia Garzon Mora  
Sarah-Tanja Hess  
Emil Horobet  
Nidhi Kaihnsa  
Elzbieta Polak  
Hamid Rahkooy  
Bernhard Reinke  
Andrea Rosana  
Felix Rydell  
Pierpaola Santarsiero  
Victoria Schleis  
Svala Sverrisdóttir  
Ettore Teixeira Turatti  
Máté László Telek  
Angelica Marcela Torres Bustos  
Beihui Yuan

Their feedback was tremendously important for the development of this book. Without their input this work would not have been possible.



# Contents

|          |   |    |
|----------|---|----|
| <b>1</b> | <b>Historical Snapshot</b>                | 1  |
| 1.1      | Polars                                    | 2  |
| 1.2      | Foci                                      | 5  |
| 1.3      | Envelopes                                 | 9  |
| <b>2</b> | <b>Critical Equations</b>                 | 13 |
| 2.1      | Euclidean Distance Degree                 | 14 |
| 2.2      | Low Rank Matrix Approximation             | 18 |
| 2.3      | Invitation to Polar Degrees               | 21 |
| <b>3</b> | <b>Computations</b>                       | 25 |
| 3.1      | Gröbner Bases                             | 26 |
| 3.2      | The Parameter Continuation Theorem        | 30 |
| 3.3      | Polynomial Homotopy Continuation          | 32 |
| <b>4</b> | <b>Polar Degrees</b>                      | 37 |
| 4.1      | Polar Varieties                           | 38 |
| 4.2      | Projective Duality                        | 40 |
| 4.3      | Chern Classes                             | 42 |
| <b>5</b> | <b>Wasserstein Distance</b>               | 45 |
| 5.1      | Polyhedral Norms                          | 46 |
| 5.2      | Optimal Transport and Independence Models | 48 |
| 5.3      | Wasserstein meets Segre-Veronese          | 51 |
| <b>6</b> | <b>Curvature</b>                          | 57 |
| 6.1      | Plane Curves                              | 58 |
| 6.2      | Algebraic Varieties                       | 64 |
| 6.3      | Volumes of Tubular Neighborhoods          | 66 |
| <b>7</b> | <b>Medial Axis and Reach</b>              | 69 |
| 7.1      | Bottlenecks                               | 71 |
| 7.2      | Offset Hypersurfaces                      | 73 |
| 7.3      | Offset Discriminant                       | 75 |

|   |     |
|---|-----|
| <b>8 Voronoi Cells</b>                          | 79  |
| 8.1 Voronoi Basics                              | 80  |
| 8.2 Computing Algebraic Boundaries              | 81  |
| 8.3 Formulas from Algebraic Geometry            | 85  |
| 8.4 Voronoi meets Eckhart-Young                 | 87  |
| <b>9 Condition Numbers</b>                      | 91  |
| 9.1 Errors in Numerical Computations            | 92  |
| 9.2 Matrix Inversion and Eckhart-Young          | 95  |
| 9.3 Condition Number Theorems                   | 98  |
| 9.4 Distance to the Discriminant                | 102 |
| <b>10 Machine Learning</b>                      | 105 |
| 10.1 Neural Networks                            | 106 |
| 10.2 Convolutional Networks                     | 110 |
| 10.3 Learning Algebraic Varieties               | 113 |
| <b>11 Maximum Likelihood</b>                    | 117 |
| 11.1 Kullback-Leibler Divergence                | 118 |
| 11.2 Maximum Likelihood Degree                  | 119 |
| 11.3 Scattering Equations                       | 122 |
| 11.4 Gaussian Models                            | 124 |
| <b>12 Tensors</b>                               | 127 |
| 12.1 Tensors and their Rank                     | 128 |
| 12.2 Eigenvectors and Singular Vectors          | 132 |
| 12.3 Volumes of Rank-One Varieties              | 135 |
| <b>13 Computer Vision</b>                       | 139 |
| 13.1 Multiview Varieties                        | 140 |
| 13.2 Grassmann Tensors                          | 144 |
| 13.3 3D Reconstruction from Unknown Cameras     | 146 |
| <b>14 Volumes</b>                               | 149 |
| 14.1 Calculus and Beyond                        | 150 |
| 14.2 D-Modules                                  | 151 |
| 14.3 Lasserre's Method                          | 158 |
| <b>15 Sampling</b>                              | 165 |
| 15.1 Computing the Homology from Finite Samples | 166 |
| 15.2 Sampling with Density Guarantees           | 167 |
| 15.3 Sampling from Probability Distributions    | 169 |
| <b>References</b>                               | 179 |

# **Chapter 1**

## **Historical Snapshot**

Throughout this book, we will encounter many instances of the interplay of metric concepts with algebraic objects. In classical texts such as Salmon's book [161], metric properties of algebraic varieties were essential. This includes the curvature of algebraic curves and computing their arc lengths and areas using integral calculus. Conversely, many curves of interest were defined in terms of distances or angular conditions. This chapter provides an introduction to this classical approach to algebraic curves.

In the following chapters we will develop a modern view on *metric algebraic geometry*. This involves the consideration of algebraic structures in distance minimization (Chapter 2), optimal transport (Chapter 5), machine learning (Chapter 10), or computer vision (Chapter 13). Consequently, we will often use the word *model* when talking about algebraic varieties, because in these settings algebraic varieties serve as an abstract mathematical model for data. In particular, we pose the *variety hypothesis*: information in data can be described by algebraic relations.

Our starting point is typically a *real algebraic variety*  $X \subset \mathbb{R}^n$ . To utilize methods from algebraic geometry, we pass to the complex Zariski closure of  $X$  in  $n$ -dimensional complex affine space  $\mathbb{C}^n$  or complex projective space  $\mathbb{P}^n$ . In most cases, to avoid heavy notation we will use the same symbol for both  $X$  and its (projective or affine) complexification.

The metric in metric algebraic geometry enters in form of a distance function. In most of the following Chapters the distance is induced by the standard Euclidean structure on  $\mathbb{R}^n$ . Over the real numbers, we use the Euclidean inner product. For  $\mathbf{p}, \mathbf{q} \in \mathbb{R}^n$  it is defined as

$$\langle \mathbf{p}, \mathbf{q} \rangle := \mathbf{p}^T \mathbf{q} = \sum_{i=1}^n p_i q_i.$$

This induces the Euclidean norm

$$\|\mathbf{p}\| := \sqrt{\langle \mathbf{p}, \mathbf{p} \rangle}.$$

The Euclidean distance function

$$d(\mathbf{p}, \mathbf{q}) := \|\mathbf{p} - \mathbf{q}\|,$$

turns  $\mathbb{R}^n$  into a metric space. The usual metric in  $\mathbb{C}^n$  is given by *Hermitian inner product*  $\langle \mathbf{p}, \mathbf{q} \rangle_{\mathbb{C}} = \mathbf{p}^* \mathbf{q}$ . However, complex conjugation is not an algebraic operation, and this is why we avoid using it. Instead, throughout this book we consider the *algebraic extension* of the Euclidean inner product and write  $\langle \mathbf{p}, \mathbf{q} \rangle = \mathbf{p}^T \mathbf{q}$  also when  $\mathbf{p}, \mathbf{q} \in \mathbb{C}^n$  are complex points. This yields a non-degenerate bilinear form on  $\mathbb{C}^n$ , which we can use to identify  $\mathbb{C}^n$  with its dual space. This way, we can consider an algebraic variety and its dual within the same ambient space. We emphasize that over the complex numbers  $\langle \cdot, \cdot \rangle$  is *not* positive definite, and hence does not induce a metric. For instance,  $\langle (1, i), (1, i) \rangle = 0$ . Nevertheless, using the algebraic extension of the Euclidean inner product we will be able to study metric problems in  $\mathbb{R}^n$  using methods from complex algebraic geometry. In some chapters we will use other metrics, but this will be spelled out explicitly. For instance, in Chapter 5 we care about distance functions induced by a norms whose unit ball is a polyhedron. However, the standard choice of metric is the Euclidean metric above.

## 1.1 Polars

Our first historical snapshot is on polars of plane curves. This will lead us to metric definitions of curves.

For a fixed line  $L$  in the real plane with a distinguished point  $\mathbf{o} \in L$ , we can choose a positive and negative direction and define a signed Euclidean distance on  $L$ . More concretely, we fix a unit vector  $\mathbf{v}$  through  $\mathbf{o}$  spanning  $L$  and define

$$\overline{\mathbf{op}} := \lambda \in \mathbb{R}, \quad \text{where } \mathbf{p} = \mathbf{o} + \lambda \mathbf{v} \in L.$$

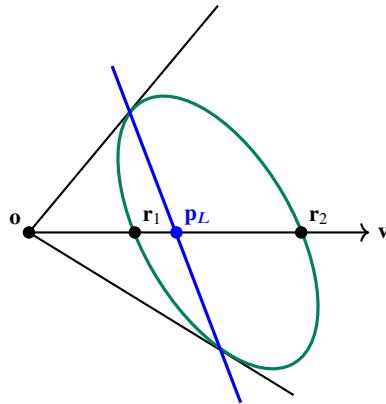


Fig. 1.1: A conic and its polar line (blue) with respect to the point  $\mathbf{o}$ .

In particular, this definition depends on the line  $L$ , the reference point  $\mathbf{o}$ , and the chosen direction given by the vector  $\mathbf{v}$ . The unsigned distance is the Euclidean distance

$$|\overline{\mathbf{op}}| = \|\mathbf{o} - \mathbf{p}\|.$$

According to Salmon [161, §56], the following theorem was first proven by Roger Cotes (1682-1716) in his *Harmonia Mensurarum*:

**Theorem 1.1 (Cotes [53])** Consider an algebraic plane curve  $C$  of degree  $d$  and a fixed point  $\mathbf{o}$  in the plane. For any line  $L$  through the point  $\mathbf{o}$  intersecting the curve  $C$  in  $d$  points  $\mathbf{r}_1, \dots, \mathbf{r}_d$ , we denote by  $\mathbf{p}_L$  the point on  $L$  whose signed distance to  $\mathbf{o}$  satisfies

$$\frac{d}{\overline{\mathbf{op}_L}} = \frac{1}{\overline{\mathbf{or}_1}} + \frac{1}{\overline{\mathbf{or}_2}} + \dots + \frac{1}{\overline{\mathbf{or}_d}}. \quad (1.1)$$

Define

$$K := \bigcup_{L \text{ is a line through } \mathbf{o}} \{\mathbf{p}_L \mid \mathbf{p}_L \text{ satisfies Equation (1.1)}\}$$

Then  $K$  is a straight line.

Salmon called the line  $K$  in Theorem 1.1 the *polar line* of the curve  $C$  and the point  $\mathbf{o}$ . The polar line is defined in metric terms.

**Example 1.2** Consider a conic  $C$  and a point  $\mathbf{o}$  outside of  $C$ . Then the polar line is spanned by the two points on  $C$  whose tangent line contains  $\mathbf{o}$ ; see Figure 1.1.  $\diamond$

The relation (1.1) means that the signed distance from  $\mathbf{o}$  to  $\mathbf{p}_L$  is the harmonic mean of the signed distances from  $\mathbf{o}$  to the intersection points  $\mathbf{r}_i$ . We rewrite this as

$$\sum_{i=1}^d \left( \frac{1}{\overline{\mathbf{op}_L}} - \frac{1}{\overline{\mathbf{or}_i}} \right) = 0. \quad (1.2)$$

In case the point  $\mathbf{o}$  is at infinity, the point  $\mathbf{p}_L$  becomes the average of the points  $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_d$ . This is proved in the next corollary.

**Corollary 1.3** Let  $C$  be a plane curve of degree  $d$  and  $\mathcal{L}$  a pencil of parallel lines. The locus of all points

$$\mathbf{p}_L := \frac{1}{d} \sum_{i=1}^d \mathbf{r}_i, \quad \text{where } L \cap C = \{\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_d\}, \quad (1.3)$$

that are the averages of the intersection points of  $L \in \mathcal{L}$  with  $C$ , is a straight line.

**Proof** We start by fixing a line  $L \in \mathcal{L}$ . Since  $\overline{\mathbf{o}\mathbf{r}_i} - \overline{\mathbf{o}\mathbf{p}_L} = \overline{\mathbf{o}\mathbf{r}_i} + \overline{\mathbf{p}_L\mathbf{o}} = \overline{\mathbf{p}_L\mathbf{r}_i}$ , Equation (1.2) is equivalent to

$$\sum_{i=1}^d \frac{\overline{\mathbf{p}_L\mathbf{r}_i}}{\overline{\mathbf{o}\mathbf{r}_i}} = 0. \quad (1.4)$$

In the latter calculation, we kept the same choice of direction on the line  $L$ , but changed the reference point from  $\mathbf{o}$  to  $\mathbf{p}_L$ . To investigate what happens in the limit when  $\mathbf{o}$  goes to infinity, we fix a point  $\mathbf{o}_0$  on the line  $L$  such that  $\lambda_i := \overline{\mathbf{o}_0\mathbf{r}_i} > 0$ . All the points on  $L$  beyond  $\mathbf{o}_0$  we can express as  $\mathbf{o}_t := \mathbf{o}_0 - t\mathbf{v}$ , where  $t \geq 0$  and  $\mathbf{v}$  is the vector defining  $L$  and its direction. Then, we have  $\overline{\mathbf{o}_t\mathbf{r}_i} = t + \lambda_i$ . For every fixed  $t \geq 0$ , we write  $\mathbf{p}_{L,t}$  for the point  $\mathbf{p}_L$  that satisfies (1.4) for  $\mathbf{o}_t$ . Multiplying (1.4) with  $t$ , we obtain

$$\sum_{i=1}^d \frac{t}{t + \lambda_i} \cdot \overline{\mathbf{p}_{L,t}\mathbf{r}_i} = 0.$$

In the limit  $t \rightarrow \infty$ , the point  $\mathbf{p}_{L,t}$  converges to a point  $\mathbf{p}_L$  such that  $\sum_{i=1}^d \overline{\mathbf{p}_L\mathbf{r}_i} = 0$ , which means that  $\mathbf{p}_L$  is the average of the points  $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_d$ . Hence, for  $\mathbf{o}$  at infinity, Equation (1.1) describes the average point  $\mathbf{p}_L$  in (1.3) and Theorem 1.1 specializes to Corollary 1.3.  $\square$

Salmon attributes Corollary 1.3 to Newton who called the resulting straight line the *diameter* of the curve  $C$  corresponding to the parallel-lines pencil  $\mathcal{L}$  [161, §51] (Cotes and Newton knew each other. In fact, Cotes edited the second edition of Newton's Principia before its publication). In contemporary applied mathematics, this result has been extended to higher-dimensional varieties and is known as the *trace test* in numerical algebraic geometry [166, Chapter 15.5].

**Remark 1.4** Another theorem on distances of a point to intersection points between a curve and lines that was first given by Newton in his *Enumeratio Linearum Tertii Ordinis* is the following: For a plane curve  $C$  of degree  $d$ , a point  $\mathbf{o}$  in the plane, and two distinct lines passing through  $\mathbf{o}$ , consider the ratio

$$\frac{\overline{\mathbf{o}\mathbf{r}_1} \cdot \overline{\mathbf{o}\mathbf{r}_2} \cdots \overline{\mathbf{o}\mathbf{r}_d}}{\overline{\mathbf{o}\mathbf{s}_1} \cdot \overline{\mathbf{o}\mathbf{s}_2} \cdots \overline{\mathbf{o}\mathbf{s}_d}}, \quad (1.5)$$

where  $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_d$  and  $\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_d$  are the intersection points of the two lines with  $C$ . Then the ratio (1.5) is invariant under translations of the point  $\mathbf{o}$  and the two intersecting lines [161, §46].

In [161, §57], Salmon generalizes the construction of polar lines to polar curves of higher order. Using the same notation as in Theorem 1.1, he shows that the locus of points  $\mathbf{p}_L$  satisfying

$$\sum_{1 \leq i < j \leq d} \left( \frac{1}{\overline{\mathbf{o}\mathbf{p}_L}} - \frac{1}{\overline{\mathbf{o}\mathbf{r}_i}} \right) \left( \frac{1}{\overline{\mathbf{o}\mathbf{p}_L}} - \frac{1}{\overline{\mathbf{o}\mathbf{r}_j}} \right) = 0$$

instead of (1.1) is a conic, called the *polar conic* of the curve  $C$  and the point  $\mathbf{o}$ . If  $\mathbf{o}$  is at infinity, the polar conic is also referred to as the *diametral conic*. More generally, the locus of points  $\mathbf{p}_L$  satisfying

$$\sum_{i_1 < \dots < i_k} \left( \frac{1}{\overline{\mathbf{op}_L}} - \frac{1}{\overline{\mathbf{or}_{i_1}}} \right) \left( \frac{1}{\overline{\mathbf{op}_L}} - \frac{1}{\overline{\mathbf{or}_{i_2}}} \right) \dots \left( \frac{1}{\overline{\mathbf{op}_L}} - \frac{1}{\overline{\mathbf{or}_{i_k}}} \right) = 0$$

is the *polar curve* of order  $k$  associated with the curve  $C$  and the point  $\mathbf{o}$  [161, §58]. If  $\mathbf{o}$  is at infinity, that polar curve is called the *curvilinear diameter* of order  $k$ .

Although polar curves have a metric origin, in modern algebro-geometric literature polar curves are typically defined using a metric-free approach; see Chapter 4. The polar curves can be expressed without involving distances as follows. We write  $f(x, y, z) = 0$  for the defining equation of the curve  $C$  in homogeneous coordinates. For  $\mathbf{o} = (a : b : c)$ , we define the operator

$$\Delta_{\mathbf{o}} := a \frac{\partial}{\partial x} + b \frac{\partial}{\partial y} + c \frac{\partial}{\partial z}.$$

Then, letting  $d := \deg C$ , the polynomial  $\Delta_{\mathbf{o}}^{d-k} f$  is the defining equation of the polar curve of order  $k$ . This is proved by Salmon in [161, §63].

By exploiting symmetry in Taylor series, Salmon also shows the following beautiful duality between polar curves that “may be written at pleasure” [161, §63]. Setting  $\mathbf{p} = (x : y : z)$ , Salmon’s result states:

$$\frac{1}{(d-k)!} \Delta_{\mathbf{o}}^{d-k} f(x, y, z) = \frac{1}{k!} \Delta_{\mathbf{p}}^k f(a, b, c),$$

In particular, the polar curve of order  $k$  that is associated with the degree- $d$  curve  $C$  and the point  $\mathbf{o}$  is the locus of all points  $\mathbf{p}$  such that the polar curve of order  $d - k$  associated with  $C$  and  $\mathbf{p}$  passes through  $\mathbf{o}$ .

## 1.2 Foci

“we believe that it will be found that every point which has any special relation to any curve will be found either to be a singular point of the curve, or a focus of it” [161, §125]

An ellipse is the locus of points in a plane whose sum of distances to two fixed points in the plane is constant. The two points are the *foci* of the ellipse. In 1832, Plücker generalized the definition of foci to arbitrary plane curves as follows. Consider a circle in the plane. If we embed the plane into two dimensional projective space  $\mathbb{P}^2$  by sending  $(x, y) \in \mathbb{C}^2$  to  $(x : y : 1)$  every circle passes through the two *circular points at infinity*  $(1 : i : 0)$  and  $(1 : -i : 0)$ . In fact, passing through both circular points completely characterizes circles.

**Lemma 1.5** *A quadric is a circle if and only if it passes through the two circular points.*

**Proof** Let  $f = (x - az)^2 + (y - bz)^2 - rz^2$  be a circle. Then,  $f(1 : \pm i : 0) = 1^2 + (\pm i)^2 = 0$ . This shows one direction. For the other direction, let  $f = a_1x^2 + a_2xy + a_3y^2 + a_4xz + a_5yz + a_6z^2$  be a real quadric. If  $f$  passes through the circular points, then

$$0 = f(1 : i : 0) = a_1 + a_2i - a_3, \quad 0 = f(1 : -i : 0) = a_1 - a_2i - a_3.$$

This implies  $a_1 = a_3$  and  $a_2 = 0$ . Hence,  $f$  is a circle.  $\square$

The following definition goes back to Plücker [151].

**Definition 1.6** Consider a plane curve  $C$ . A point  $\mathbf{f}$  in the plane is a *focus* of the curve  $C$  if both lines spanned by the point  $\mathbf{f}$  and the circular points at infinity are tangent to the curve  $C$ .

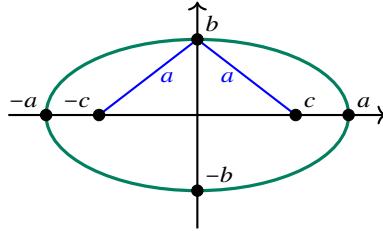


Fig. 1.2: The real foci of an ellipse with width  $2a$  and height  $2b$  ( $a \geq b$ ) and its axes aligned with the  $x$ - and  $y$ -axis of the real affine plane are  $(\pm\sqrt{a^2 - b^2}, 0)$ .

**Example 1.7** We determine the foci (in the sense of Plücker's Definition 1.6) of an ellipse. Since the condition that a line in  $\mathbb{P}^2$  passes through the point  $(1 : \pm i : 0)$  is invariant under translations and rotations of the real affine plane  $\{z \neq 0\} \subseteq \mathbb{P}^2$ , we may translate and rotate the ellipse such that its two defining foci (whose sum of distances is constant along the ellipse) become  $(c, 0)$  and  $(-c, 0)$  in the real affine plane (with  $c \geq 0$ ). Now the ellipse is the locus of points satisfying

$$\sqrt{(x - c)^2 + y^2} + \sqrt{(x + c)^2 + y^2} = 2a,$$

for some constant  $a > 0$ . From this, we see that the two points  $(\pm a, 0)$  lie on the ellipse. Moreover, the ellipse intersects the  $y$ -axis at  $(0, \pm b)$  such that  $b^2 = a^2 - c^2$ ; see Figure 1.2. Hence, the width and height of the ellipse are  $2a$  and  $2b$ , respectively, (with  $a \geq b > 0$ ) and its defining equation can alternatively be written as

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

If the ellipse is not a circle (i.e.,  $a > b$ ), then there are two tangent lines to the ellipse that pass through the circular point at infinity  $(1 : i : 0)$ , namely the tangent lines at  $\mathbf{p}_+ := (a^2 : ib^2 : c)$  and  $\mathbf{p}_- := (-a^2 : -ib^2 : c)$ . The tangent lines at the complex conjugates  $\bar{\mathbf{p}}_+$  and  $\bar{\mathbf{p}}_-$  of the latter two points pass through the other circular point at infinity  $(1 : -i : 0)$ . The foci à la Plücker are the four points of intersection of the two tangent lines through  $(1 : i : 0)$  with the two tangent lines through  $(1 : -i : 0)$ . Since the tangent lines at  $\mathbf{p}_\pm$  and  $\bar{\mathbf{p}}_\pm$  are a complex conjugated pair, they meet at a real point, namely  $(\pm c : 0 : 1)$ . This shows that the real foci we used to define the ellipse are indeed foci in the sense of Plücker's Definition 1.6. The other two foci are obtained by intersecting the tangent line at  $\mathbf{p}_\pm$  with the tangent line at  $\bar{\mathbf{p}}_\mp$ . They are the imaginary points  $(0 : \mp ic : 1)$ .

If the ellipse is a circle (i.e.,  $a = b$ ), it passes through the circular points at infinity. The tangent lines at those points are the only ones that pass through the circular points at infinity. They are given by  $x \pm iy = 0$  and intersect at the origin  $(0 : 0 : 1)$ . The four foci of a general ellipse come together to a single point, namely the center of the circle.

In general, the number of foci of an algebraic plane curve depends on its *class*, that is the degree of its dual curve. The *dual projective plane* is the set of lines in the original projective plane  $\mathbb{P}^2$ . The *dual curve*  $C^\vee$  of a plane curve  $C \subseteq \mathbb{P}^2$  is the Zariski closure in the dual projective plane of the set of tangent lines at regular points of  $C$ . Hence, the degree of the dual curve  $C^\vee$  (equivalently, the class of  $C$ ) is the number of tangent lines to  $C$  that pass through a generic point in the plane  $\mathbb{P}^2$ . Using the non-degenerate bilinear form  $\langle \cdot, \cdot \rangle$  coming from the Euclidean inner product, we can view the dual curve as a curve in  $\mathbb{P}^2$ . More specifically,

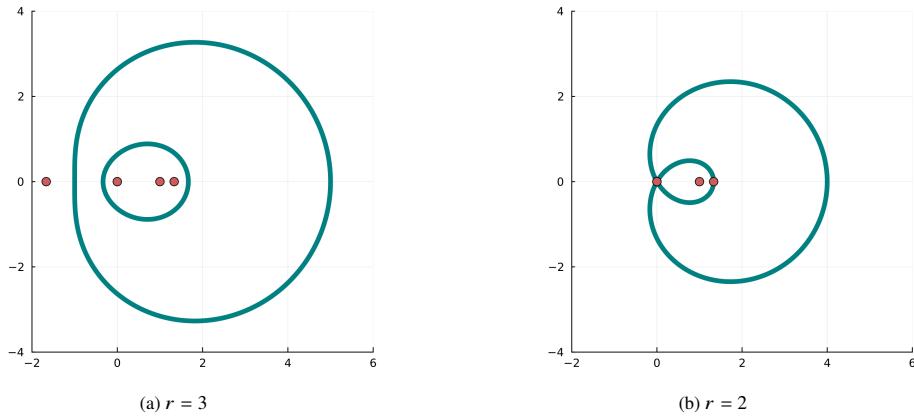


Fig. 1.3: Cartesian ovals with defining foci  $\mathbf{f}_1 = (0, 0)$ ,  $\mathbf{f}_2 = (1, 0)$  and weight  $s = 2$  are described by the polynomial equation  $(-3(x^2 + y^2) + 8x + r^2 - 4)^2 - 4r^2(x^2 + y^2) = 0$ . In both cases, the focus at  $(-4/3, 0)$  has order three. In the right picture, the focus at  $(0, 0)$  has order two. This gives 6 real foci in total, respectively.

$$C^\vee = \overline{\{\mathbf{y} \mid \text{there is a regular point } \mathbf{x} \in C \text{ with } \langle \mathbf{y}, \mathbf{v} \rangle = 0 \text{ for all } \mathbf{v} \in T_{\mathbf{x}}C\}} \subset \mathbb{P}^2. \quad (1.6)$$

This means  $C^\vee$  is the Zariski closure of the union of normal lines to  $C$  (when orthogonality is measured by  $\langle \cdot, \cdot \rangle$ ). If  $C$  is defined by the polynomial  $f$  and  $\mathbf{x} \in C$ , the gradient  $\nabla f(\mathbf{x}) = (\partial f / \partial x_1(\mathbf{x}), \partial f / \partial x_2(\mathbf{x}))$  spans the normal line of  $C$  at  $\mathbf{x}$ . Consequently,

$$C^\vee = \overline{\{\nabla f(\mathbf{x}) \mid \mathbf{x} \text{ is a regular point of } C\}}.$$

Suppose now that the curve  $C$  has class  $m$ . Then, through each of the two circular points at infinity, there are  $m$  tangent lines to the curve  $C$ . There are  $m^2$  intersection points of these two sets of  $m$  lines. Furthermore, when the curve  $C$  is real the intersection points of conjugate pairs of tangent lines are also real. This leads to the following result; see [161, §125].

**Proposition 1.8** *A curve  $C$  of class  $m$  has  $m^2$  complex foci (counted with multiplicity). When the curve  $C$  is real, exactly  $m$  foci are real*

In Example 1.7 we have worked out the foci of an ellipse. In the remainder of this section, we want to compute the foci of additional classes of curves. There are several constructions that generalize ellipses in an obvious way. For instance, an *n-ellipse* is the locus of points in a plane whose sum of distances to  $n$ -fixed points in the plane is constant. *n*-ellipses were studied by Tschirnhaus in 1686 [174] and Maxwell in 1846 [131]. In general, *n*-ellipses are not algebraic, but semi-algebraic. In fact, they are special cases of spectrahedra [139]. Here, we focus on an alternative generalization. Instead of considering the sum of distances to two foci, we consider a *weighted sum* or *weighted product*. This leads to so-called Cartesian and the Cassini ovals.

A *Cartesian oval* (named after Descartes who first studied them in his 1637 *La Géométrie* for their application to optics) is the locus of points in a plane whose weighted sum of distances to two fixed points is constant, i.e., it is the locus of points  $\mathbf{p}$  that satisfy

$$\|\mathbf{p} - \mathbf{f}_1\| + s\|\mathbf{p} - \mathbf{f}_2\| = r,$$

where  $\mathbf{f}_1$  and  $\mathbf{f}_2$  are fixed points in the plane and  $s, r$  are fixed constants. In general, that Cartesian oval is not an algebraic curve, but it does satisfy a quartic equation. In fact, the real vanishing locus of that quartic polynomial is two nested ovals; see Figure 1.3a. Among the four possible equations

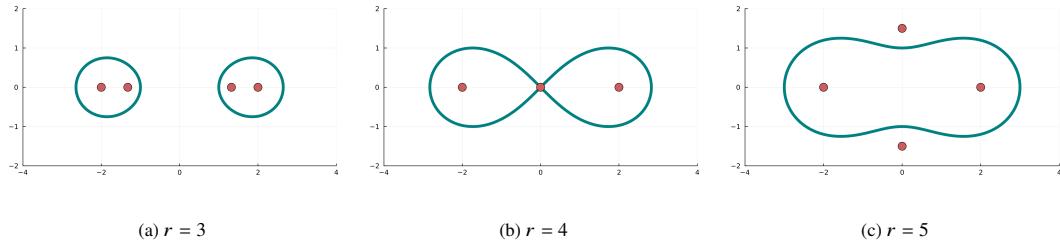


Fig. 1.4: Cassini ovals with defining foci  $\mathbf{f}_1 = (-2, 0)$ ,  $\mathbf{f}_2 = (2, 0)$  and their remaining foci. The outer foci on the  $x$ -axis all have multiplicity 3. The middle focus in picture (b) has multiplicity 2. The equation of the algebraic curves is in (1.8).

$$\|\mathbf{p} - \mathbf{f}_1\| \pm s\|\mathbf{p} - \mathbf{f}_2\| = \pm r, \quad (1.7)$$

exactly two have real solutions and those describe the ovals. Salmon shows that a quartic curve is a Cartesian oval if and only if it has cusps at the two circular points at infinity [161, §129]. By Plücker's formula [161, §72], it follows that the class of such a quartic is six (except in degenerate cases). Salmon determines the six real foci of Cartesian oval quartic curves [161, §129] (see also Basset [12, §273]): Three of them form a triple focus which is located at the intersection of the cusps' tangent lines. The remaining three foci lie on a straight line. Two of them are the points  $f_1$  and  $f_2$  that define the curve in (1.7). It was already observed by Chasles that in fact any two of the three single foci can be used to define the Cartesian oval by an equation of the form (1.7) [47, Note XXI]. If two of the three single foci come together, the Cartesian oval degenerates to a *Limaçon of Pascal*; see Figure 1.3b. Salmon further observes another interesting metric property of foci: Whenever a line meets a Cartesian oval in four points, the sum of their four distances from any of the three single foci is constant [161, §218].

A *Cassini oval* (named after Cassini who studied them in 1693 [42]) is the locus of points in a plane whose product of distances to two fixed points  $\mathbf{f}_1$  and  $\mathbf{f}_2$  is constant. This is an algebraic curve defined by the real quartic polynomial

$$\|\mathbf{p} - \mathbf{f}_1\|^2 \cdot \|\mathbf{p} - \mathbf{f}_2\|^2 = r^2$$

for some constant  $r$ . The circular points at infinity are double points on each Cassini oval, and since these are the only singularities (except in degenerate cases), the class of a Cassini oval is eight in general [161, §219]. Basset [12, §247] explains that the two pairs of complex conjugated tangent lines at the two nodes intersect at  $\mathbf{f}_1$  and  $\mathbf{f}_2$ , respectively. Hence, those points are foci in the sense of Plücker's Definition 1.6. In fact, Basset shows that each of them is a triple focus (the reason being that the nodal tangents are stationary). To describe the remaining two real foci, we translate and rotate (as in Example 1.7) such that  $\mathbf{f}_1 = (c, 0)$  and  $\mathbf{f}_2 = (-c, 0)$ . Then, the Cassini oval is defined by

$$((x - c)^2 + y^2)((x + c)^2 + y^2) = r^2. \quad (1.8)$$

If  $r < c^2$ , the real locus is two ovals. Otherwise, the real locus is connected, where the degenerate case  $r = c^2$  is the *lemniscate of Bernoulli*; see Figure 1.4. In the case of two ovals, the remaining two real foci also lie on the  $x$ -axis; they are  $(\pm \frac{1}{c} \sqrt{c^4 - r^2}, 0)$ . If  $r > c^2$ , the two foci are  $(0, \pm \frac{1}{c} \sqrt{r^2 - c^4})$  and lie on the  $y$ -axis. In the degenerate case  $r = c^2$ , they become a double focus at the origin.

### 1.3 Envelopes

Given a one-dimensional algebraic family of lines in the plane, its *envelope* is a curve such that each of the given lines is tangent to the curve. Equivalently, in the dual projective plane, the family of lines is an algebraic curve  $\mathcal{L}$  and its dual curve  $\mathcal{L}^\vee$  is the envelope.

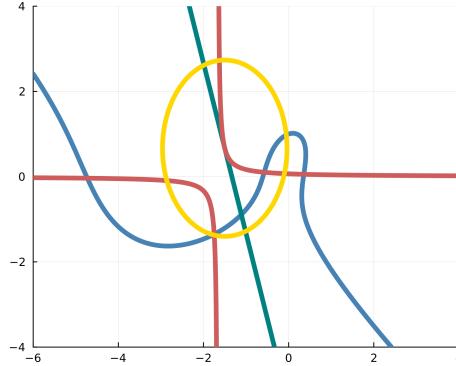


Fig. 1.5: Cubic curve  $x^3 + y^3 + 5x^2 - 2xy + x - 1 = 0$  (blue) and the envelope  $9xy + 15y - 1 = 0$  (red) of its diameters. The diameter (green) and diametral conic (yellow) are associated with the point  $(2 : 1 : 0)$  at infinity.

**Example 1.9** The diameters of a plane curve  $C$  form a one-dimensional family of lines. For a cubic curve  $C$ , the envelope of that family is the locus consisting of the centers of the diametral conics of  $C$  [161, §160]. For instance, for the cubic curve defined by  $x^3 + y^3 + 5x^2 - 2xy + x - 1 = 0$ , the diameter associated with the point  $(a : b : 0)$  at infinity is the line given by  $3a^2x + 3b^2y + 5a^2 - 2ab = 0$ . The family of all diameters is the curve  $\mathcal{L}$  in the dual plane that is defined by  $25x^2 - 4xy - 30x + 9 = 0$ . Its dual curve  $\mathcal{L}^\vee$ , that is the envelope of the diameters, is  $9xy + 15y - 1 = 0$ . The diametral conic associated with  $(a : b : 0)$  is  $3a\left(x + \frac{5a-b}{3a}\right)^2 + 3b\left(y - \frac{a}{3b}\right)^2 + a - \frac{(5a-b)^2}{3a} - \frac{a^2}{3b} = 0$  and so its center  $\left(\frac{-5a+b}{3a}, \frac{a}{3b}\right)$  lies on the envelope. In fact, the tangent line of the envelope at that point is the diameter associated with  $(a : b : 0)$ ; see Figure 1.5.

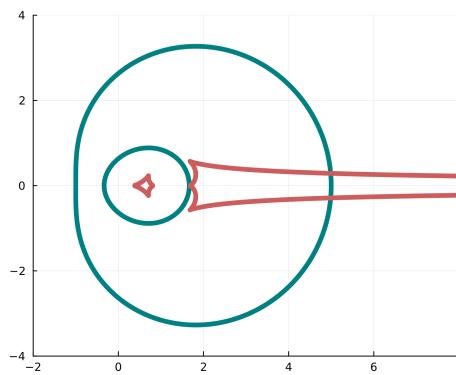


Fig. 1.6: The Cartesian oval from Figure 1.3a together with its evolute.

Special instances of envelopes that are defined by metric properties are evolutes and caustics, both of which we will discuss in the remainder of this chapter. The *evolute* of a plane curve  $C$  is the envelope of its normals (i.e., the lines orthogonal to its tangent lines). Equivalently, the evolute is the locus of the centers of curvature (see Proposition 6.2). The study of evolutes goes back to Apollonius (ca. 200 BC) [177]. A recent discussion of evolutes can be found in [148]. The degree and class of the evolute of a general smooth curve of degree  $d$  are  $3d(d - 1)$  (see Corollary 6.9) and  $d^2$  [161, §116], respectively. Moreover, its evolute has  $\frac{d}{2}(3d - 5)(3d^2 - d - 6)$  double points,  $3d(2d - 3)$  cusps, and no other singularities. For instance, the evolute of the Cartesian oval in Figure 1.3a is illustrated in Figure 1.6. Its equation is

$$\begin{aligned} & 102036672x^{10}y^2 + 433655856x^8y^4 + 833407380x^6y^6 + 917059401x^4y^8 + 558336726x^2y^{10} + 143065521y^{12} \\ & - 884317824x^9y^2 - 3106029888x^7y^4 - 4898885832x^5y^6 - 4008450240x^3y^8 - 1331276472xy^{10} + 3251316672x^8y^2 \\ & + 9515584512x^6y^4 + 12088352844x^4y^6 + 6432939486x^2y^8 + 620191890y^{10} - 40310784x^9 - 6758774784x^7y^2 \\ & - 16647933888x^5y^4 - 15962551632x^3y^6 - 4237194240xy^8 + 342641664x^8 + 9145229184x^6y^2 + 18728830368x^4y^4 \\ & + 11743648812x^2y^6 + 961612425y^8 - 1239556608x^7 - 9234062208x^5y^2 - 14497919136x^3y^4 - 4640798304xy^6 \\ & + 2495722752x^6 + 8064660672x^4y^2 + 8003654064x^2y^4 + 835700656y^6 - 3071831040x^5 - 6288399360x^3y^2 \\ & - 2974296960xy^4 + 2390342400x^4 + 3772699200x^2y^2 + 540271200y^4 - 1173312000x^3 - 1396800000xy^2 \\ & + 349920000x^2 + 228000000y^2 - 57600000x + 4000000 = 0. \end{aligned}$$

We will see in Proposition 6.5 that the finite cusps of the evolute correspond to the points of critical curvature of the original curve  $C$ . Salmon computes the length of an arc of the evolute as “the difference of the radii of curvature at its extremities” [161, §115]. The converse operation of computing the evolute is finding an *involute*, that is, for a given plane curve  $C$ , find any curve whose evolute is  $C$ . In contrast to evolutes, involutes of an algebraic curve are typically not unique (they are parallel curves / offset curves; see Section 7.2) and they might not be algebraic. For instance, the involute of a circle is a transcendental curve [161, §235]. Nevertheless, the foci of a plane curve are also foci of its evolute and involute [161, §127].

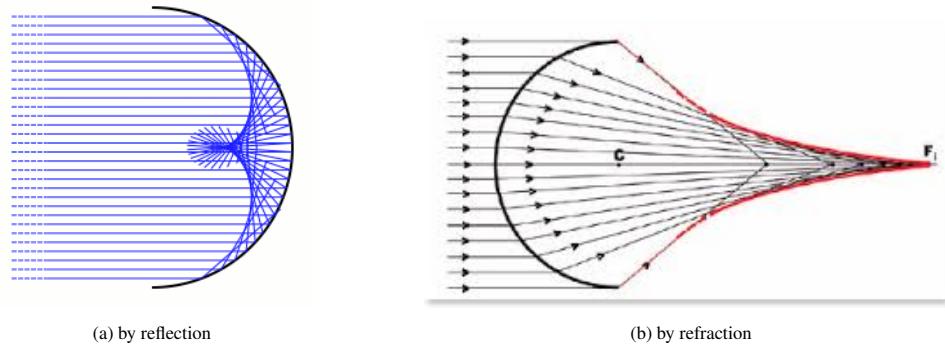


Fig. 1.7: Caustics of a circle with light source at infinity.

Caustics of plane curves come in two flavors. Let us imagine that a fixed point in the plane emits light. The light rays get reflected at each point of a given plane curve. The *caustic by reflection* is the envelope of the family of reflected rays. Similarly, the *caustic by refraction* is the envelope of the family of refracted rays. Figure 1.7 shows caustics of a circle when the light source is at infinity. Those curves can be commonly observed in real life, e.g., when the sun shines on a round glass.

**To appear: Example by Felix Rydell:** The caustics by refractions of conics are the evolutes of Cartesian ovals (original result due to M. Quetelet). In the limit, when conics degenerate to double lines, one sees that caustics by refraction of lines are evolutes of conics.



## **Chapter 2**

### **Critical Equations**

We consider a model  $X$  that is given as the zero set in  $\mathbb{R}^n$  of a collection  $\{f_1, \dots, f_k\}$  of nonlinear polynomials in  $n$  unknowns  $x_1, \dots, x_n$ . Thus,  $X$  is a real algebraic variety. In order to apply algebraic methods, it is preferable to work with the complex algebraic variety defined by the same polynomials. We will use the same symbol for this:  $X \subset \mathbb{C}^n$ . We assume that  $X$  is irreducible, that  $I_X = \langle f_1, \dots, f_k \rangle$  is its prime ideal, and that the set of nonsingular real points is Zariski dense in  $X$ . The  $k \times n$  Jacobian matrix  $\mathcal{J} = (\partial f_i / \partial x_j)$  has rank at most  $c$  at any point  $\mathbf{x} \in X$ , where  $c = \text{codim}(X)$ . The point  $\mathbf{x}$  is *nonsingular* on  $X$  if the rank is exactly  $c$ . The variety  $X$  is called *smooth* if all its points are nonsingular. Elaborations on these hypotheses are found in many text books, including [135, Chapter 2].

The following optimization problem arises in many applications. Given a data point  $\mathbf{u} \in \mathbb{R}^n \setminus X$ , compute the distance to the model  $X$ . Thus, we seek a point  $\mathbf{x}^*$  in  $X$  that is closest to  $\mathbf{u}$ . The answer depends on the chosen metric. We focus on the case when the metric is represented by a polynomial and  $\mathbf{x}^*$  is a smooth point on  $X$ . The optimal point  $\mathbf{x}^*$  is a solution to the *critical equations*. In optimization, these are also known as first-order conditions or KKT equations, and they arise from introducing Lagrange multipliers. We seek to compute all complex solutions to the critical equations. The set of these *critical points* is typically finite, and it includes all local maxima, all local minima and all saddle points.

## 2.1 Euclidean Distance Degree

We begin by discussing the *Euclidean distance (ED) problem*, which is as follows:

$$\text{minimize } \|\mathbf{x} - \mathbf{u}\|^2 = \sum_{i=1}^n (x_i - u_i)^2 \text{ subject to } \mathbf{x} = (x_1, \dots, x_n) \in X. \quad (2.1)$$

This optimization problem is defined over the real locus of  $X$ . We study it by the following strategy: we first derive the critical equations for (2.1), and then study them over the complex variety  $X$ . This way, we can interpret the optimization problem (2.1) as a problem of solving polynomial equations over the complex numbers.

The *augmented Jacobian matrix*  $\mathcal{AJ}$  is the  $(k+1) \times n$  matrix which is obtained by placing the row vector  $(x_1 - u_1, \dots, x_n - u_n)$  atop the Jacobian matrix  $\mathcal{J}$ . We form the ideal generated by its  $(c+1) \times (c+1)$  minors, we add the ideal of the model  $I_X$ , and we then saturate that sum by the ideal of  $c \times c$  minors of  $\mathcal{J}$ . See [62, Eqn. (2.1)]. The result is the *critical ideal*  $C_{X,\mathbf{u}}$  of the model  $X$  with respect to the given data point  $\mathbf{u}$ .

**Example 2.1 (Plane curves)** Let  $X$  be the plane curve defined by a polynomial  $f(x_1, x_2)$  in two unknowns. We wish to compute the Euclidean distance from  $X$  to a given point  $\mathbf{u} = (u_1, u_2) \in \mathbb{R}^2$ . To this end, we form the augmented Jacobian matrix. This matrix is square of size  $2 \times 2$ :

$$\mathcal{AJ} = \begin{pmatrix} x_1 - u_1 & x_2 - u_2 \\ \partial f / \partial x_1 & \partial f / \partial x_2 \end{pmatrix} \quad (2.2)$$

The critical ideal is obtained from  $f$  and the determinant of  $\mathcal{AJ}$  by performing a saturation step:

$$C_{X,\mathbf{u}} = \langle f, \det(\mathcal{AJ}) \rangle : \langle \partial f / \partial x_1, \partial f / \partial x_2 \rangle^\infty. \quad (2.3)$$

The ideal  $C_{X,\mathbf{u}}$  lives in  $\mathbb{R}[x_1, x_2]$ . Frequently, the coefficients of  $f$  and the coordinates of  $\mathbf{u}$  are rational numbers, and in this case we can perform the computation purely symbolically in  $\mathbb{Q}[x_1, x_2]$ . The saturation step in (2.3) removes points that are singular on the curve  $X = V(f)$ . If  $X$  is smooth then saturation is unnecessary, and we simply have  $C_{X,\mathbf{u}} = \langle f, \det(\mathcal{AJ}) \rangle$ . On the other hand, if  $X$  has singular points,

we must saturate. For a concrete example take the cardioid

$$f = (x_1^2 + x_2^2 + x_3)^2 - (x_1^2 + x_2^2), \quad (2.4)$$

and fix a random point  $\mathbf{u} = (u_1, u_2)$ . See [62, Example 1.1]. The ideal  $\langle f, \det(\mathcal{AJ}) \rangle$  is the intersection of  $C_{X,\mathbf{u}}$  and an  $\langle x_1, x_2 \rangle$ -primary ideal of multiplicity 3. The critical ideal  $C_{X,\mathbf{u}}$  has three distinct complex zeros. We can express their coordinates in radicals in the given numbers  $u_1, u_2$ . The following code computes the degree of  $C_{X,\mathbf{u}}$  for  $\mathbf{u} = (2, 1)$  using the software Macaulay2 [79].

```
R = QQ[x1, x2];
u1 = 2; u2 = 1;
f = (x1^2+x2^2+x3)^2 - (x1^2+x2^2)
AJ = matrix {{x1-u1, x2-u2}, {diff(x1, f), diff(x2, f)}}
I = ideal {f, det AJ}
C = saturate(I, ideal {diff(x1, f), diff(x2, f)})
degree C
```

◊

*Remark 2.2* Saturation is discussed in more detail in Chapter 3, in particular the definition of the saturation of an ideal is in Equation 3.1.

The variety  $V(C_{X,\mathbf{u}})$  is the set of complex critical points of (2.1). For random data  $\mathbf{u}$ , this variety is a finite subset of  $\mathbb{C}^n$ , and it contains the optimal solution  $\mathbf{x}^*$ , provided the latter is attained at a smooth point of  $X$ . It was proved in [62] that the number of critical points, i.e. the cardinality of the variety  $V(C_{X,\mathbf{u}})$ , is independent of  $\mathbf{u}$ , if we assume that the data point  $\mathbf{u}$  is sufficiently general. See also Example 3.20 in the next chapter.

**Definition 2.3** The *Euclidean Distance Degree* (EDD) of the variety  $X$  is the number of points in the variety  $V(C_{X,\mathbf{u}})$ . We write this as

$$\text{EDdegree}(X) := \#V(C_{X,\mathbf{u}}).$$

In Example 2.1 we examined a plane curve of degree 4 whose ED degree equals 3. The ED degree of a variety  $X$  measures the difficulty of solving the ED problem (2.1) using exact algebraic methods. The ED degree is an important complexity measure in metric algebraic geometry.

**Example 2.4 (Space curves)** Fix  $n = 3$  and let  $X$  be the curve in  $\mathbb{R}^3$  defined by two general polynomials  $f_1$  and  $f_2$  of degrees  $d_1$  and  $d_2$  in three unknowns  $x_1, x_2, x_3$ . The augmented Jacobian matrix is

$$\mathcal{AJ} = \begin{pmatrix} x_1 - u_1 & x_2 - u_2 & x_3 - u_3 \\ \partial f_1 / \partial x_1 & \partial f_1 / \partial x_2 & \partial f_1 / \partial x_3 \\ \partial f_2 / \partial x_1 & \partial f_2 / \partial x_2 & \partial f_2 / \partial x_3 \end{pmatrix}. \quad (2.5)$$

Fix a general data vector  $\mathbf{u} \in \mathbb{R}^3$ . Then the critical ideal equals  $C_{X,\mathbf{u}} = \langle f_1, f_2, \det(\mathcal{AJ}) \rangle$ . Hence the set of critical points is the intersection of three surfaces. These surfaces have degrees  $d_1, d_2$  and  $d_1 + d_2 - 1$ . By Bézout's Theorem [135, Theorem 2.16], the expected number of solutions is the product of these degrees. Hence the ED degree of the curve  $X$  equals  $d_1 d_2 (d_1 + d_2 - 1)$ .

The same formula can be derived from a formula for general curves in terms of algebraic geometry data. Let  $X$  be a general smooth curve of degree  $d$  and genus  $g$  in any ambient space  $\mathbb{R}^n$ . By [62, Corollary 5.9], we have  $\text{EDdegree}(X) = 3d + 2g - 2$ . The above curve in 3-space has degree  $d = d_1 d_2$  and genus  $g = d_1^2 d_2 / 2 + d_1 d_2^2 / 2 - 2d_1 d_2 + 1$ . We conclude that

$$\text{EDdegree}(X) = 3d + 2g - 2 = d_1 d_2 (d_1 + d_2 - 1).$$

This formula also covers the case of plane curves (cf. Example 2.1). Namely, if we set  $d_1 = d$  and  $d_2 = 1$  then we see that a general plane curve  $X$  of degree  $d$  has  $\text{EDdegree}(X) = d^2$ . In particular, a general plane quartic has ED degree 16. However, that number can drop a lot for curves that are special. For the cardioid in (2.4) the ED degree drops from 16 to 3.  $\diamond$

Here is a general upper bound on the ED degree in terms of the given polynomials.

**Proposition 2.5** *Let  $X$  be a variety of codimension  $c$  in  $\mathbb{R}^n$  whose ideal  $I_X$  is generated by polynomials  $f_1, f_2, \dots, f_c, \dots, f_k$  of degrees  $d_1 \geq d_2 \geq \dots \geq d_c \geq \dots \geq d_k$ . Then*

$$\text{EDdegree}(X) \leq d_1 d_2 \cdots d_c \cdot \sum_{i_1+i_2+\dots+i_c \leq n-c} (d_1 - 1)^{i_1} (d_2 - 1)^{i_2} \cdots (d_c - 1)^{i_c}. \quad (2.6)$$

*Equality holds when  $X$  is a generic complete intersection of codimension  $c$  (hence  $c = k$ ).*

**Proof** This appears in [62, Proposition 2.6]. We can derive it as follows. Bézout's Theorem ensures that the degree of the variety  $X$  is at most  $d_1 d_2 \cdots d_c$ . The entries in the  $i$ th row of the matrix  $\mathcal{AJ}$  are polynomials of degrees  $d_{i-1} - 1$ . The degree of the variety of  $(c+1) \times (c+1)$  minors of  $\mathcal{AJ}$  is at most the sum in (2.6). This follows from the Giambelli–Thom–Porteous formula, which expresses the degree of a determinantal variety in terms of symmetric functions. The intersection of that determinantal variety with  $X$  is our set of critical points, and the cardinality of that set is bounded by the product of the two degrees. Generically, that intersection is a complete intersection and the inequality (2.6) is attained.  $\square$

Formulas or a priori bounds for the ED degree are important when studying exact solutions to the optimization problem (2.1). The paradigm is to compute all complex critical points, by either symbolic or numerical methods, and to then extract one's favorite real solutions among these. This leads, for instance, to all local minima in (2.1). The ED degree is an upper bound on the number of real critical points, but this bound is generally not tight.

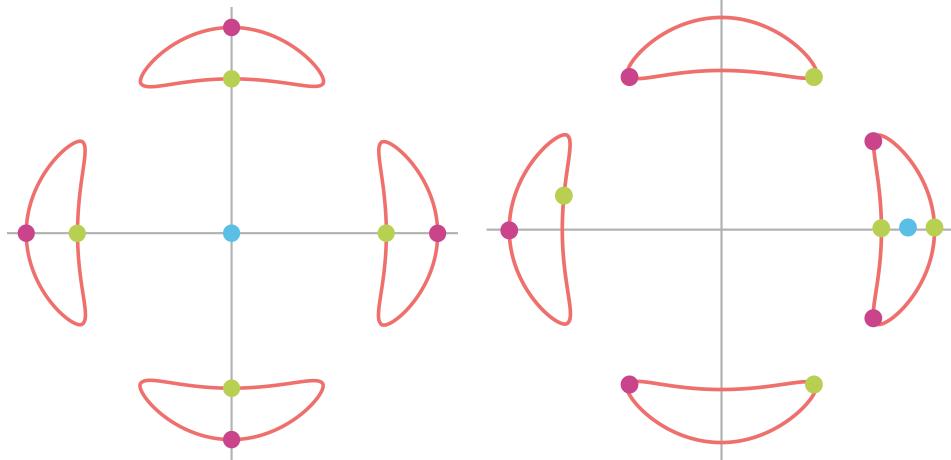


Fig. 2.1: ED problems on the Trott curve: configurations of eight (left) or ten (right) critical points. Data points are blue, local minimal are green, and local maxima are purple. The coordinates of the critical points are computed by solving the critical equations in (2.7).

**Example 2.6** Consider the case  $n = 2, c = 1, d_1 = 4$  in Proposition 2.5, where  $X$  is a generic quartic curve in the plane  $\mathbb{R}^2$ . The number of complex critical points is  $\text{EDdegree}(X) = 16$ . But, they cannot be all real. For an illustration, consider the *Trott curve*  $X = V(f)$ , given by

$$f = 144(x_1^4 + x_2^4) - 225(x_1^2 + x_2^2) + 350x_1^2x_2^2 + 81.$$

This curve is shown in Figure 2.1. For general data  $\mathbf{u} = (u_1, u_2)$  in  $\mathbb{R}^2$ , the critical equations

$$f = \frac{\partial f}{\partial x_2}(x_1 - u_1) - \frac{\partial f}{\partial x_1}(x_2 - u_2) = 0. \quad (2.7)$$

have distinct 16 complex solutions, and these are all critical points in  $X$ . Since the Trott curve is smooth, the saturation step in (2.3) is not needed when computing the ideal  $C_{X,\mathbf{u}}$ .

The ED degree 16 is an upper bound for the number of real critical points of the optimization problem (2.1) for any data point  $\mathbf{u}$ . The actual number of real critical points depends heavily on the specific location of  $\mathbf{u}$ . For data  $\mathbf{u}$  near the origin, eight of the 16 points in  $V(C_{X,\mathbf{u}})$  are real. For  $\mathbf{u} = (\frac{7}{8}, \frac{1}{100})$ , which is inside the rightmost oval, there are 10 real critical points. The two scenarios are shown in Figure 2.1. Local minima are green, while local maxima are purple. Finally, consider  $\mathbf{u} = (2, \frac{1}{100})$ , which lies to the right of the rightmost oval. Here, the number of real critical points is 12.  $\diamond$

In general, our task is to compute the complex zeros of the critical ideal  $C_{X,\mathbf{u}}$ . Algorithms for this computation can be either symbolic or numerical. Symbolic methods usually rest on the construction of a Gröbner basis, to be followed by a floating point computation to extract the solutions. In recent years, numerical methods have become popular. These are based on homotopy continuation. We explain how homotopy continuation works in the Chapter 3, called *computations*.

**Example 2.7** Let us revisit the cardioid from Example 2.1. For  $\mathbf{u} = (2, 1)$  we compute the three critical points using the Julia [21] software HomotopyContinuation.jl [29].

```
using HomotopyContinuation, LinearAlgebra
@var x1 x2
u1 = 2; u2 = 1;
f = (x1^2+x2^2+x2)^2 - (x1^2+x2^2)
AJ = [[x1 - u1; x2 - u2] differentiate(f, [x1;x2])]
F = [f; det(AJ)]
solve(F)
```

$\diamond$

We next illustrate current capabilities.

**Example 2.8** Suppose  $X$  is defined by  $c = k = 3$  random polynomials in  $n = 7$  variables, for a range of degrees  $d_1, d_2, d_3$ . The table below lists the ED degree in each case, and the times used by HomotopyContinuation.jl to compute and certify all critical points in  $\mathbb{C}^7$ .

| $d_1 \ d_2 \ d_3$ | 3 2 2 | 3 3 2 | 3 3 3 | 4 2 2 | 4 3 2 | 4 3 3 | 4 4 2 | 4 4 3 |
|-------------------|-------|-------|-------|-------|-------|-------|-------|-------|
| EDdegree          | 1188  | 3618  | 9477  | 4176  | 10152 | 23220 | 23392 | 49872 |
| Solving (sec)     | 3.849 | 21.06 | 61.51 | 31.51 | 103.5 | 280.0 | 351.5 | 859.3 |
| Certifying (sec)  | 0.390 | 1.549 | 4.653 | 2.762 | 7.591 | 17.16 | 21.65 | 50.07 |

Here we represent  $C_{X,\mathbf{u}}$  by a system of 10 equations in 10 variables. In addition to the three equations  $f_1 = f_2 = f_3 = 0$  in  $x_1, \dots, x_7$ , we take the seven equations  $(1, y_1, y_2, y_3) \cdot \mathcal{AJ} = 0$ . Here  $y_1, y_2, y_3$  are new variables. These additional equations ensure that the  $4 \times 7$  matrix  $\mathcal{AJ}$  has rank  $\leq 3$ . This formulation avoids the listing of all  $\binom{7}{4} = 35$  maximal minors. It is the preferred representation of determinantal varieties in the setting of numerical algebraic geometry.

The timings above refer to computing all complex solutions to the system of 10 equations in 10 variables. They include the certification step [33] that proves correctness and completeness. These computations were performed using HomotopyContinuation.jl v2.5.6 on a 16 GB MacBook Pro with an Intel Core

i7 processor working at 2.6 GHz. They suggest that our critical equations can be solved fast and reliably, with proof of correctness, when the ED degree is less than 50000. When the ED degree exceeds 50000, success with numerical path tracking will depend on the specific structure of the family. A key player on the geometric side is the discriminant of the problem. If that is well-behaved, then even larger ED degrees are feasible. A successful application to a physics problem is reported in [171, Table 1].  $\diamond$

We now introduce an alternative version of the Euclidean Distance Degree by using a different Euclidean metric. We consider *weight vector*  $\Lambda = (\lambda_1, \dots, \lambda_n) \in \mathbb{R}^n$  whose entries are positive real numbers. The *weighted Euclidean metric* is then

$$\|\mathbf{x}\|_{\Lambda}^2 = \sum_{i=1}^n \lambda_i \mathbf{x}_i^2$$

If  $\Lambda$  is the all-one vector, we obtain the usual Euclidean metric. We can then consider the analogue of the optimization problem (2.1):

$$\text{minimize } \|\mathbf{x} - \mathbf{u}\|_{\Lambda}^2 = \sum_{i=1}^n \lambda_i (x_i - u_i)^2 \text{ subject to } \mathbf{x} = (x_1, \dots, x_n) \in X. \quad (2.8)$$

Let  $\mathbf{u} \in \mathbb{R}^n$ . Similar to above we obtain the  $\Lambda$ -augmented Jacobian matrix  $\mathcal{AJ}_{\Lambda} \in \mathbb{R}^{(k+1) \times n}$  by placing the row vector  $(\lambda_1(x_1 - u_1), \dots, \lambda_1(x_n - u_n))$  atop the Jacobian matrix  $\mathcal{J}$ . The critical ideal of the optimization problem (2.8) is then  $C_{X,\mathbf{u}}^{\Lambda} = I : J^{\infty}$ , where  $I$  is generated by the  $(c+1) \times (c+1)$  minors of  $\mathcal{AJ}_{\Lambda}$  together with the generators of  $I_X$ , and  $J$  is the ideal of  $c \times c$  minors of  $\mathcal{J}$ . This leads to the following definition.

**Definition 2.9** Let  $X \subset \mathbb{R}^n$  be a real variety,  $\Lambda = (\lambda_1, \dots, \lambda_n)$  have positive entries, and  $\mathbf{u} \in \mathbb{R}^n$  be a general point. The  $\Lambda$ -weighted Euclidean Distance Degree of the variety  $X$  is the number of points in the variety  $V(C_{X,\mathbf{u}}^{\Lambda})$ . For general  $\Lambda$  this number is constant, and we call it the generic Euclidean Distance Degree of  $X$ , denoted

$$\text{EDdegree}_{\text{gen}}(X) := \#V(C_{X,\mathbf{u}}^{\Lambda}).$$

## 2.2 Low Rank Matrix Approximation

A specific instance of the ED problem (2.1) arises when considering varieties of matrices of low rank that are constrained to have a special structure. Sometimes these matrices are flattenings of tensors. This version of the problem was studied in the article [144], which focuses on Hankel matrices, Sylvester matrices and generic subspaces of matrices, and which uses a weighted version of the Euclidean metric. In this section we offer a brief introduction to this special case of our general ED problem.

Our point of departure is the following low-rank approximation problem for rectangular matrices:

$$\text{minimize } \|A - U\|^2 = \sum_{i=1}^m \sum_{j=1}^n (a_{ij} - u_{ij})^2 \quad \text{subject to } \text{rank}(A) \leq r. \quad (2.9)$$

In this problem, we are given a real *data matrix*  $U = (u_{ij})$  of format  $m \times n$ , and we wish to find a real matrix  $A = (a_{ij})$  of rank at most  $r$  that is closest to  $U$ . The norm for matrices is sometimes called *Frobenius norm*. We can write this as  $\|A\|^2 = \text{Trace}(A^T A)$ . Indeed, if  $\mathbf{a}_1, \dots, \mathbf{a}_n$  denote the columns of  $A$ , the diagonal entries of  $A^T A$  are  $\langle \mathbf{a}_i, \mathbf{a}_i \rangle$  for  $i = 1, \dots, n$ , so that

$$\text{Trace}(A^T A) = \sum_{i=1}^n \langle \mathbf{a}_i, \mathbf{a}_i \rangle = \sum_{i=1}^n \sum_{j=1}^m a_{ij}^2 = \|A\|^2$$

The solution to (2.9) is given by the *singular value decomposition*

$$U = T_1 \cdot \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_m) \cdot T_2.$$

Here  $T_1, T_2$  are orthogonal matrices, and  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_m$  are the singular values of  $U$ . The following well-known theorem from numerical linear algebra concerns the variety of  $m \times n$  matrices of rank  $\leq r$ .

**Theorem 2.10 (Eckart-Young)** *The closest matrix of rank  $\leq r$  to the given matrix  $U$  equals*

$$U^* = T_1 \cdot \text{diag}(\sigma_1, \dots, \sigma_r, 0, \dots, 0) \cdot T_2. \quad (2.10)$$

This is the unique local minimum. All complex critical points are real. They are found by substituting zeros for  $m - r$  of the entries of  $\text{diag}(\sigma_1, \dots, \sigma_m)$ . Hence,  $\text{EDdegree}(X) = \binom{m}{r}$ .

We give a complete proof of the Eckart-Young theorem in Section 9.2.

In *structured low-rank approximation* we are given a linear subspace  $\mathcal{L} \subset \mathbb{R}^{m \times n}$ , and a data matrix  $U \in \mathcal{L}$ , and we wish to solve the restricted problem:

$$\text{minimize } \|A - U\|^2 = \sum_{i=1}^m \sum_{j=1}^n (a_{ij} - u_{ij})^2 \text{ subject to } A \in \mathcal{L} \text{ and } \text{rank}(A) \leq r. \quad (2.11)$$

A best-case scenario would be the following: if  $U$  lies in  $\mathcal{L}$  then so does the SVD solution  $U^*$  in (2.10). This happens for some linear subspaces  $\mathcal{L}$ , including symmetric and circulant matrices. However, most subspaces  $\mathcal{L}$  of  $\mathbb{R}^{m \times n}$  do not enjoy this property, and finding the global optimum in (2.11) can be quite difficult. The article [144] studies this optimization problem for both generic and special subspaces  $\mathcal{L}$ . It rests on [62] and uses tools from algebraic geometry.

We study low-rank and structured low-rank approximation using the weighted Euclidean metric. As before, our primary task is to compute the number of complex critical points of (2.11). Thus, we seek to find the Euclidean distance degree (ED degree) of the determinantal variety

$$\mathcal{L}_{\leq r} := \{A \in \mathcal{L} \mid \text{rank}(A) \leq r\}.$$

This variety is always regarded as a subvariety of the matrix space  $\mathbb{R}^{m \times n}$ .

We use the  $\Lambda$ -weighted Euclidean distance coming from  $\mathbb{R}^{m \times n}$ . We write  $\text{EDdegree}_{\Lambda}(\mathcal{L}_{\leq r})$  for the  $\Lambda$ -weighted Euclidean distance degree of the variety  $\mathcal{L}_{\leq r}$  (see Definition 2.9). The importance of the weights  $\Lambda$  is highlighted in [62, Example 3.2], for the seemingly harmless situation when  $\mathcal{L}$  is the space of all symmetric matrices in  $\mathbb{R}^{n \times n}$ .

Of special interest are the the usual Euclidean Distance Degree  $\text{EDdegree}(\mathcal{L}_{\leq r})$ , when  $\Lambda = \mathbf{1}$  is the all-one matrix, and the generic ED degree  $\text{EDdegree}_{\text{gen}}(\mathcal{L}_{\leq r})$ , when the weight matrix  $\Lambda$  is generic. The generic ED degree is given by a formula that rests on intersection theory. See [62, Theorem 7.7] and Theorem 2.14 below. Indeed, choosing the weights  $\lambda_{ij}$  to be generic ensures that the projective closure of  $\mathcal{L}_{\leq r}$  has transversal intersection with the isotropic quadric  $\{A \in \mathbb{P}^{mn-1} \mid \sum_{i=1}^m \sum_{j=1}^n \lambda_{ij} a_{ij}^2 = 0\}$ .

We next present two examples that illustrate the concepts above. These can then also serve as examples for Theorem 2.14 below, as seen by the Macaulay2 [79] calculation in Example 2.22.

**Example 2.11** Let  $m = n = 3$  and  $\mathcal{L} \subset \mathbb{R}^{3 \times 3}$  the 5-dimensional space of Hankel matrices:

$$A = \begin{bmatrix} a_0 & a_1 & a_2 \\ a_1 & a_2 & a_3 \\ a_2 & a_3 & a_4 \end{bmatrix}, \quad U = \begin{bmatrix} u_0 & u_1 & u_2 \\ u_1 & u_2 & u_3 \\ u_2 & u_3 & u_4 \end{bmatrix} \quad \text{and} \quad \Lambda = \begin{bmatrix} \lambda_0 & \lambda_1 & \lambda_2 \\ \lambda_1 & \lambda_2 & \lambda_3 \\ \lambda_2 & \lambda_3 & \lambda_4 \end{bmatrix}.$$

Our goal in (2.11) is to solve the following constrained optimization problem for  $r = 1, 2$ :

$$\begin{aligned} \text{minimize } & \lambda_0(a_0 - u_0)^2 + 2\lambda_1(a_1 - u_1)^2 + 3\lambda_2(a_2 - u_2)^2 + 2\lambda_3(a_3 - u_3)^2 + \lambda_4(a_4 - u_4)^2 \\ \text{subject to } & \text{rank}(A) \leq r. \end{aligned}$$

This can be rephrased as an unconstrained optimization problem. For instance, for rank  $r = 1$ , we get a one-to-one parametrization of  $\mathcal{L}_{\leq 1}$  by setting  $a_i = st^i$ . Our optimization problem is as follows:

$$\text{minimize } \lambda_0(s - u_0)^2 + 2\lambda_1(st - u_1)^2 + 3\lambda_2(st^2 - u_2)^2 + 2\lambda_3(st^3 - u_3)^2 + \lambda_4(st^4 - u_4)^2.$$

The ED degree is the number of critical points with  $t \neq 0$ . We consider three weight matrices:

$$\mathbf{1} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \quad \Omega = \begin{bmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/2 \\ 1/3 & 1/2 & 1 \end{bmatrix}, \quad \Theta = \begin{bmatrix} 1 & 2 & 2 \\ 2 & 2 & 2 \\ 2 & 2 & 1 \end{bmatrix}.$$

Here  $\Omega$  gives the usual Euclidean metric when  $\mathcal{L}$  is identified with  $\mathbb{R}^5$ . The last weight matrix  $\Theta$  arises from identifying  $\mathcal{L}$  with the space of symmetric  $2 \times 2 \times 2 \times 2$ -tensors; see Chapter 12 for a study of metrics in the space of tensors. We compute

$$\begin{aligned} \text{EDdegree}_\mathbf{1}(\mathcal{L}_{\leq 1}) &= 6, & \text{EDdegree}_\Omega(\mathcal{L}_{\leq 1}) &= 10, & \text{EDdegree}_\Theta(\mathcal{L}_{\leq 1}) &= 4, \\ \text{EDdegree}_\mathbf{1}(\mathcal{L}_{\leq 2}) &= 9, & \text{EDdegree}_\Omega(\mathcal{L}_{\leq 2}) &= 13, & \text{EDdegree}_\Theta(\mathcal{L}_{\leq 2}) &= 7. \end{aligned}$$

In both cases,  $\Omega$  exhibits the generic behavior, so we have  $\text{EDdegree}_{\text{gen}}(\mathcal{L}_{\leq r}) = \text{EDdegree}_\Omega(\mathcal{L}_{\leq r})$ . We refer to [144, Sections 3 and 4] for larger Hankel matrices and formulas for their ED degrees.  $\diamond$

*Remark 2.12* The distance given by the weight matrix  $\Theta$  in the previous example yields the *Bombieri-Weyl distance* that plays an important role in Chapter 9.

**Example 2.13** Let  $m = n = 3, r = 1$  but now take  $\mathcal{L} = \mathbb{R}^{3 \times 3}$ . Thus, we are considering the weighted rank-one approximation problem for  $3 \times 3$ -matrices. We know from [62, Example 7.10] that  $\text{EDdegree}_{\text{gen}}(\mathcal{L}_{\leq 1}) = 39$ . We take a circulant data matrix and a circulant weight matrix:

$$U = \begin{bmatrix} -59 & 11 & 59 \\ 11 & -59 & -59 \\ 59 & -59 & 11 \end{bmatrix} \quad \text{and} \quad \Lambda = \begin{bmatrix} 9 & 6 & 1 \\ 6 & 1 & 9 \\ 1 & 9 & 6 \end{bmatrix}.$$

This instance has 39 complex critical points. Of these, 19 are real, and 7 are local minima:

$$\begin{aligned} & \begin{bmatrix} 0.0826 & 2.7921 & -1.5452 \\ 2.7921 & 94.3235 & -52.2007 \\ -1.5452 & -52.2007 & 28.8890 \end{bmatrix}, \begin{bmatrix} -52.2007 & 28.8890 & -1.5452 \\ 2.7921 & -1.5452 & 0.0826 \\ 28.8890 & -1.5452 & -52.2007 \end{bmatrix}, \begin{bmatrix} -52.2007 & 2.7921 & 94.3235 \\ 2.7921 & 94.3235 & -52.2007 \\ -1.5452 & 0.0826 & 2.7921 \end{bmatrix}, \\ & \begin{bmatrix} -29.8794 & 36.2165 & -27.2599 \\ -32.7508 & 39.6968 & -29.8794 \\ 39.6968 & -48.1160 & 36.2165 \end{bmatrix}, \begin{bmatrix} -48.1160 & 36.2165 & 39.6968 \\ 36.2165 & -27.2599 & -29.8794 \\ 39.6968 & -29.8794 & -32.7508 \end{bmatrix}, \begin{bmatrix} -29.8794 & -32.7508 & 39.6968 \\ 36.2165 & 39.6968 & -48.1160 \\ -27.2599 & -29.8794 & 36.2165 \end{bmatrix}, \\ & \begin{bmatrix} -25.375 & -25.375 & -25.375 \\ -25.375 & -25.375 & -25.375 \\ -25.375 & -25.375 & -25.375 \end{bmatrix}. \end{aligned}$$

The first three are the global minima in our ED distance problem. The last matrix is the local minimum where the objective function has the largest value: note that each entry equals  $-203/8$ . The entries of the first six matrices are algebraic numbers of degree 10 over  $\mathbb{Q}$ . For instance, the two upper left entries 0.0826 and  $-48.1160$  are among the four real roots of the irreducible polynomial

$$\begin{aligned} & 164466028468224x^{10} + 2785648335954688x^9 + 1602205386689376672x^8 + 728536260028875412x^7 \\ & -2198728936046680414272x^6 - 14854532690380098143152x^5 + 2688673091228371095762316x^4 \\ & +44612094455115888622678587x^3 - 41350080445712457319337106x^2 \\ & +27039129499043116889674775x - 1977632463563766878765625. \end{aligned}$$

Here, the critical ideal in  $\mathbb{Q}[x_{11}, x_{12}, \dots, x_{33}]$  is not prime. It is the intersection of six maximal ideals. Their degrees over  $\mathbb{Q}$  are 1, 2, 6, 10, 10, 10. The sum of these numbers equals  $39 = \text{EDdegree}_{\text{gen}}(\mathcal{L}_{\leq 1})$ .  $\diamond$

Explicit formulas are derived in [144, Section 3] for  $\text{EDdegree}_{\text{gen}}(\mathcal{L}_{\leq r})$  when  $\mathcal{L}$  is a generic subspace of  $\mathbb{R}^{m \times n}$ . This covers the four cases that arise by pairing affine subspaces or linear subspaces with either unit weights or generic weights. One important feature of determinantal varieties is they are not complete intersections. Their ED degrees are much smaller than suggested by the upper bound in Proposition 2.5.

## 2.3 Invitation to Polar Degrees

We have introduced the ED degree of an algebraic variety  $X$  as a complexity measure for the ED problem in (2.1). The number 39 in the previous example served as an illustration on how the ED degree controls the number of critical points. But a deeper understanding is needed. In this section, we develop the algebro-geometric roots of the ED degree, which will then yield more advanced algorithms for finding it.

**Theorem 2.14** *If the given variety  $X$  meets both the hyperplane at infinity and the isotropic quadric transversally, then  $\text{EDdegree}(X)$  equals the sum of the polar degrees of the projective closure of  $X$ .*

This theorem appears in [62, Proposition 6.10]. The hypothesis holds for all varieties  $X$  after a general linear change of coordinates.

We shall explain all the terms used in Theorem 2.14. First of all, the *projective closure* of our affine real variety  $X \subset \mathbb{C}^n$  is its Zariski closure in complex projective space  $\mathbb{P}^n$ , which we also denote by  $X$ . Algebraically,  $\mathbb{P}^n$  is obtained from  $\mathbb{C}^n$  by adding one homogenizing coordinate  $x_0$ . We identify the affine space  $\mathbb{C}^n$  with the open subset  $\{\mathbf{x} \in \mathbb{P}^n \mid x_0 \neq 0\}$ . Its set complement  $\{\mathbf{x} \in \mathbb{P}^n \mid x_0 = 0\} \simeq \mathbb{P}^{n-1}$  is the *hyperplane at infinity* inside  $\mathbb{P}^n$ . The hypersurface  $\{\mathbf{x} \in \mathbb{P}^{n-1} \mid \sum_{i=1}^n x_i^2 = 0\}$  is called the *isotropic quadric*. It lives in the hyperplane at infinity and it has no real points. The hypothesis in Theorem 2.14 means that the intersection of  $X$  with these two loci is reduced and has the expected dimension.

If we are given a real projective variety  $X$  in  $\mathbb{P}^n$  from the start, then we also consider the ED problem for its affine cone in  $\mathbb{R}^{n+1}$ . The data vector now equals  $\mathbf{u} = (u_0, u_1, \dots, u_n)$ , and the augmented Jacobian is redefined so as to respect the fact that all polynomials are homogeneous. The general formula for this matrix and the homogeneous critical ideal appears in [62, equation (2.7)].

For a curve  $X \subset \mathbb{P}^2$  with defining polynomial  $f(x_0, x_1, x_2)$ , we use the modified augmented Jacobian

$$\mathcal{AJ} = \begin{pmatrix} u_0 & u_1 & u_2 \\ x_0 & x_1 & x_2 \\ \partial f / \partial x_0 & \partial f / \partial x_1 & \partial f / \partial x_2 \end{pmatrix}.$$

The homogeneous critical ideal in  $\mathbb{R}[x_0, x_1, x_2]$  is computed as follows:

$$C_{X, \mathbf{u}} = \langle f, \det(\mathcal{AJ}) \rangle : (\langle \partial f / \partial x_0, \partial f / \partial x_1, \partial f / \partial x_2 \rangle \cdot (x_1^2 + x_2^2))^\infty. \quad (2.12)$$

The critical points are given by the variety  $V(C_{X,u})$  in  $\mathbb{P}^2$ , whose cardinality is  $\text{EDdegree}(X)$ . The factor  $(x_1^2 + x_2^2)$  in the saturation step (2.12) is the isotropic quadric. It is needed whenever the hypothesis of Theorem 2.14 is not satisfied. Namely, it removes any extraneous component that may arise from non-transversal intersection of the curve  $X$  with the isotropic quadric.

**Example 2.15 (Cardioid)** We consider the homogeneous version of the cardioid in Example 2.1:

$$f = (x_1^2 + x_2^2 + x_0 x_2)^2 - x_0^2 (x_1^2 + x_2^2). \quad (2.13)$$

The projective curve  $X = V(f)$  has three singular points, namely that at the origin  $V(x_1, x_2)$  in the affine plane  $\mathbb{C}^2 = \{x_0 \neq 0\}$  and the two points in the isotropic quadric  $V(x_1^2 + x_2^2)$  in  $\mathbb{P}^1 = \{x_0 = 0\}$ .

The homogeneous critical ideal  $C_{X,u}$  is generated by three cubics, and it defines seven points in  $\mathbb{P}^2$ . Hence the projective cardioid  $X$  has  $\text{EDdegree}(X) = 7$ . This is also the ED degree of the affine cardioid in (2.4) but only after a linear change of coordinates. Even a fairly modest change of coordinates can have dramatic impact. For instance, if we replace  $x_1$  by  $2x_1$  in (2.4) then the ED degree jumps from 3 to 7.  $\diamond$

We now offer a first definition of the polar degrees of a projective variety  $X \subset \mathbb{P}^n$ . Recall that points in the dual projective space  $(\mathbb{P}^n)^*$  represent hyperplanes in the primal space  $\mathbb{P}^n$ . The Euclidean bilinear form  $\langle \cdot, \cdot \rangle$  is nondegenerate. Therefore, we can use it to identify  $\mathbb{P}^n$  and its dual space, by letting  $\mathbf{h} \in \mathbb{P}^n$  represent the hyperplane  $\{\mathbf{x} \in \mathbb{P}^n \mid \langle \mathbf{h}, \mathbf{x} \rangle = h_0 x_0 + \cdots + h_n x_n = 0\}$ .

**Definition 2.16** The *conormal variety*  $N_X \subset \mathbb{P}^n \times \mathbb{P}^n$  is the Zariski closure of all pairs of points  $(\mathbf{x}, \mathbf{h})$  in  $\mathbb{P}^n \times \mathbb{P}^n$  such that  $\mathbf{x}$  is a nonsingular point of  $X$  and  $\mathbf{h}$  represents a hyperplane that is tangent to  $X$  at  $\mathbf{x}$ .

It is known that  $N_X$  has dimension  $n - 1$ , and if  $X$  is irreducible then so is  $N_X$ . The image of  $N_X$  under projection onto the second factor is the dual variety  $X^\vee$ . If  $X$  is a curve, this yields the definition of dual curve from (1.6).

The role of  $\mathbf{x}$  and  $\mathbf{h}$  in the definition of the conormal variety can be swapped. The following biduality relations [76, §I.1.3] hold:

$$N_X = N_{X^\vee} \quad \text{and} \quad (X^\vee)^\vee = X.$$

The conormal variety is an object of algebraic geometry that offers the theoretical foundations for various aspects of duality in optimization, including primal-dual algorithms.

**Example 2.17** For a plane curve  $X = V(f)$  in  $\mathbb{P}^2$ , the conormal variety  $N_X$  is a curve in  $\mathbb{P}^2 \times \mathbb{P}^2$ . Its ideal is derived from the ideal that is generated by  $f$  and the  $2 \times 2$  minors of

$$\begin{pmatrix} h_0 & h_1 & h_2 \\ \partial f / \partial x_0 & \partial f / \partial x_1 & \partial f / \partial x_2 \end{pmatrix}$$

By saturation, we remove singularities and points on the isotropic quadric, to arrive at  $C_{X,u}$ .

For instance, if  $f$  is the homogeneous cardioid in (2.13) then  $X^\vee$  is the cubic defined by

$$16h_0^3 - 27h_0h_1^2 - 24h_0^2h_2 - 15h_0h_2^2 - 2h_2^3.$$

The ideal of  $N_X$  has ten minimal generators. In addition to the above generators of bidegrees  $(4, 0)$  and  $(0, 3)$ , we find the quadric  $x_0h_0 + x_1h_1 + x_2h_2$  of bidegree  $(1, 1)$ , three cubics of bidegree  $(2, 1)$  like  $x_1^2h_1 - 3x_2^2h_1 - x_0x_1h_2 + 4x_1x_2h_2$ , and four cubics of bidegree  $(1, 2)$ .  $\diamond$

We now finally come to the polar degrees. The product of two projective spaces  $\mathbb{P}^n \times \mathbb{P}^n$  serves as the ambient space for our primal-dual approach to the ED problem. We now consider its cohomology ring:

$$H^*(\mathbb{P}^n \times \mathbb{P}^n, \mathbb{Z}) = \mathbb{Z}[s, t]/\langle s^{n+1}, t^{n+1} \rangle.$$

The class of the conormal variety  $N_X$  in this cohomology ring is a binary form of degree  $n+1 = \text{codim}(N_X)$  whose coefficients are nonnegative integers:

$$[N_X] = \delta_1(X)s^n t + \delta_2(X)s^{n-1}t^2 + \delta_3(X)s^{n-2}t^3 + \cdots + \delta_n(X)st^n. \quad (2.14)$$

**Definition 2.18** The coefficients  $\delta_i(X)$  in (2.14) are called the *polar degrees* of  $X$ .

*Remark 2.19* The polar degrees satisfy  $\delta_i(X) = \#(N_X \cap (L \times L'))$ , where  $L \subset \mathbb{P}^n$  and  $L' \subset \mathbb{P}^n$  are general linear subspaces of dimensions  $n+1-i$  and  $i$  respectively. This geometric interpretation implies that  $\delta_i(X) = 0$  for  $i < \text{codim}(X^\vee)$  and for  $i > \dim(X) + 1$ . Moreover, the first and last polar degree are the classical degrees for the dual pair of varieties:

$$\delta_i(X) = \text{degree}(X) \text{ for } i = \dim(X) + 1 \text{ and } \delta_i(X) = \text{degree}(X^\vee) \text{ for } i = \text{codim}(X^\vee). \quad (2.15)$$

**Example 2.20** Let  $X \subset \mathbb{P}^2$  be the cardioid in (2.13). The curve  $N_X \subset \mathbb{P}^2 \times \mathbb{P}^2$  has the class

$$[N_X] = \text{degree}(X^\vee) \cdot s^2t + \text{degree}(X) \cdot st^2 = 3 \cdot s^2t + 4 \cdot st^2.$$

Thus the polar degrees of the cardioid are 3 and 4. Their sum 7 is the ED degree.

**Example 2.21** Let  $X$  be a general surface of degree  $d$  in  $\mathbb{P}^3$ . Its dual  $X^\vee$  is a surface of degree  $d(d-1)^2$  in  $(\mathbb{P}^3)^\vee$ . The conormal variety  $N_X$  is a surface in  $\mathbb{P}^3 \times (\mathbb{P}^3)^\vee$ , with class

$$[N_X] = d(d-1)^2 s^3 t + d(d-1) s^2 t^2 + d s t^3.$$

The sum of the three polar degrees equals  $\text{EDdegree}(X) = d^3 - d^2 + d$ ; see Proposition 2.5.

Theorem 2.14 allows us to compute the ED degree for many interesting varieties, e.g. using Chern classes [62, Theorem 5.8]. This is relevant for applications in machine learning [35] which rest on low-rank approximation of matrices and tensors with special structure [144].

**Example 2.22 (Determinantal varieties)** Let  $X_r \subset \mathbb{P}^{m^2-1}$  be the variety of  $m \times m$  matrices  $x = (x_{ij})$  of rank  $\leq r$ . By [168], the conormal variety  $N_X$  is cut out by matrix equations:

$$N_X = \{(\mathbf{x}, \mathbf{h}) \in \mathbb{P}^{m^2-1} \times \mathbb{P}^{m^2-1} \mid \text{rank}(\mathbf{x}) \leq r, \text{rank}(\mathbf{h}) \leq m-r, \mathbf{x} \cdot \mathbf{h} = 0 \text{ and } \mathbf{h} \cdot \mathbf{x} = 0\}.$$

(Here, the product symbol  $\cdot$  means usual matrix multiplication.) This follows from Lemma 9.12 in Chapter 9. In particular, the duality relation  $(X_r)^\vee = X_{m-r}$  holds among determinantal varieties. Typing the above formula into Macaulay2, we compute the polar degrees for  $r = 1$  and  $m = 3$ :

```

QQ[x11,x12,x13,x21,x22,x23,x31,x32,x33,h11,h12,h13,h21,h22,h23,h31,h32,h33,
    Degrees=> {{1,0},{1,0},{1,0},{1,0},{1,0},{1,0},{1,0},{1,0},{1,0},
                 {0,1},{0,1},{0,1},{0,1},{0,1},{0,1},{0,1},{0,1},{0,1}}];
x = matrix {{x11,x12,x13},{x21,x22,x23},{x31,x32,x33}};
h = matrix {{h11,h12,h13},{h21,h22,h23},{h31,h32,h33}};
I = minors(2,x) + minors(3,h) + minors(1,x*h) + minors(1,h*x);
isPrime(I), codim(I), degree I
multidegree(I)

```

The code starts with the bigraded coordinate ring of  $\mathbb{P}^8 \times \mathbb{P}^8$ . It verifies that  $N_X$  has codimension 9 and that  $I$  is its prime ideal. The last command computes the polar degrees:

$$[N_X] = 3s^8t + 6s^7t^2 + 12s^6t^3 + 12s^5t^4 + 6s^4t^5. \quad (2.16)$$

After verifying (2.15), one concludes that  $\text{EDdegree}(X_1) = 3+6+12+12+6 = 39$ . Indeed, after changing coordinates, the EDdegree for  $3 \times 3$ -matrices of rank 1 equals 39. We saw this already in Example 2.13, where 39 critical points were found by a numerical computation.  $\diamond$

The primal-dual set-up of conormal varieties allows for a very elegant formulation of the critical equations. This will be presented in the next theorem. We now assume that  $X$  is an irreducible variety defined by homogeneous polynomials in  $n$  variables. Thus  $X$  is an affine cone in  $\mathbb{C}^n$ . Its dual  $Y = X^\vee$  is the affine cone over the dual of the projective variety given by  $X$ . Thus  $Y$  is also an affine cone in  $\mathbb{C}^n$ . In this setting, the conormal variety  $N_X$  is viewed as an affine variety of dimension  $n$  in  $\mathbb{C}^{2n}$ . The homogeneous ideals of these cones are precisely those discussed above.

**Theorem 2.23** *The ED problems for  $X$  and  $Y$  coincide, and we have  $\text{EDdegree}(X) = \text{EDdegree}(Y)$ . Given a general data point  $\mathbf{u} \in \mathbb{R}^n$ , the critical equations for this ED problem are:*

$$(\mathbf{x}, \mathbf{h}) \in N_X \quad \text{and} \quad \mathbf{x} + \mathbf{h} = \mathbf{u}. \quad (2.17)$$

**Proof** See [62, Theorem 5.2].  $\square$

It is instructive to verify Theorem 2.23 for Example 2.22. For any data matrix  $\mathbf{u}$  of size  $m \times m$ , the sum in (2.17) is a special decomposition of  $\mathbf{u}$ , namely  $\mathbf{x}$  of rank  $r$  plus  $\mathbf{h}$  of rank  $m - r$ . By the Eckhart-Young Theorem, it arises from zeroing out complementary singular values  $\sigma_i$  in the two matrices  $\mathbf{x}$  and  $\mathbf{h}$ .

In general, there is no free lunch, even with a simple formulation like (2.17). The difficulty lies in computing the ideal of the conormal variety  $N_X$ . However, this should be thought of as a preprocessing step, to be carried out only once per model  $X$ . If an efficient presentation of  $N_X$  is available, our task is to solve the system  $\mathbf{x} + \mathbf{h} = \mathbf{u}$  of  $n$  linear equations in  $2n$  coordinates for the  $n$ -dimensional affine variety  $N_X$ .

The discussion so far was restricted to the Euclidean norm. But, we can measure distances in  $\mathbb{R}^n$  with any other norm  $\|\cdot\|$ . Our optimization problem (2.1) extends naturally:

$$\text{minimize } \|\mathbf{x} - \mathbf{u}\| \text{ subject to } \mathbf{x} \in X. \quad (2.18)$$

The unit ball  $B = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\| \leq 1\}$  is a centrally symmetric convex body. Conversely, every centrally symmetric convex body  $B$  defines a norm, and we can paraphrase the previous optimization problem as:

$$\text{minimize } \lambda \text{ subject to } \lambda \geq 0 \text{ and } (\mathbf{u} + \lambda B) \cap X \neq \emptyset. \quad (2.19)$$

We will meet minimization problems for unit balls that are polyhedra in Chapter 5.

If the boundary of the unit ball  $B$  is smooth and algebraic then we can express the critical equations for the corresponding norm as a polynomial system. This is derived as before, but we now replace the first row of the augmented Jacobian matrix  $\mathcal{AJ}$  with the gradient of the map  $\mathbb{R}^n \rightarrow \mathbb{R}$ ,  $\mathbf{x} \mapsto \|\mathbf{x} - \mathbf{u}\|$ .

## **Chapter 3**

## **Computations**

In this chapter, we study two computational approaches to solve a system of polynomial equations. A system of  $m$  polynomial equations in  $n$  variables is a system of the form

$$F(\mathbf{x}) := \begin{pmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{pmatrix} = 0,$$

where  $f_1, \dots, f_m \in \mathbb{C}[\mathbf{x}] := \mathbb{C}[x_1, \dots, x_n]$ . If  $n = m$ , we call  $F(\mathbf{x})$  a *square system*. If  $n > m$ , we call  $F(\mathbf{x})$  *underdetermined*. If  $n < m$ , we call  $F(\mathbf{x})$  *overdetermined*. Here, we will mostly focus on square systems.

**Example 3.1** In the previous chapter, we have considered constrained optimization problems of the form  $\min_{\mathbf{x} \in \mathbb{R}^n : f(\mathbf{x})=0} g(\mathbf{x})$ , where  $f$  and  $g$  are polynomials in  $n$  variables  $\mathbf{x} = (x_1, \dots, x_n)$ . To solve this problem, one can compute the solutions of the critical equations  $f(\mathbf{x}) = \frac{\partial f}{\partial x_1} - \lambda \frac{\partial g}{\partial x_1} = \dots = \frac{\partial f}{\partial x_n} - \lambda \frac{\partial g}{\partial x_n} = 0$ . This is a square system in the  $n + 1$  variables  $(\mathbf{x}, \lambda)$ , where  $\lambda$  is Lagrange multiplier.  $\diamond$

Solving the system  $F(\mathbf{x}) = 0$  means that we compute *all* points  $\mathbf{z} = (z_1, \dots, z_n) \in \mathbb{C}^n$  such that  $F(\mathbf{z}) = 0$ . The first step is to find an appropriate *data structure* to represent a solution. In fact,  $F(\mathbf{z}) = 0$  is already *implicitly* represented by its equation. Some information can be read off from this representation. For instance, if  $F$  has rational coefficients and we know that  $F(\mathbf{z}) = \mathbf{0}$  has only finitely many complex solutions, then each coordinate  $z_i$  of  $\mathbf{z}$  is an algebraic number. On the other hand, other information like whether or there is a real solution  $\mathbf{z} \in \mathbb{R}^n$  is not directly accessible from this implicit representation.

The goal of this lecture is to discuss two data structures for representing solutions of systems of polynomial equations: the first is *Gröbner bases* and the second is *approximate zeros*.

### 3.1 Gröbner Bases

We use the notation  $\mathbf{x}^\alpha := x_1^{\alpha_1} \cdots x_n^{\alpha_n}$  for the exponent vector  $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{N}^n$ . In fact, we can identify monomials with their exponent vectors. A monomial ordering  $>$  on  $\mathbb{C}[\mathbf{x}]$  is then defined by a total order  $>$  on  $\mathbb{N}^n$  that satisfies (1) if  $\alpha > \beta$ , then  $\alpha + \gamma > \beta + \gamma$  for every  $\gamma \in \mathbb{N}^n$  and (2) every nonempty subset of  $\mathbb{N}^n$  has a smallest element under  $>$  (see, e.g., [54, Chapter 2 §2, Definition 1]).

**Example 3.2** The Lex (Lexicographic) order is defined by setting  $\alpha >_{\text{Lex}} \beta$ , if  $\alpha_j - \beta_j > 0$  for  $\alpha, \beta \in \mathbb{N}^n$ , where  $j := \min\{i \mid \alpha_i \neq \beta_i\}$  is the first index where  $\alpha$  and  $\beta$  are not equal. For instance,  $x_1^2 x_2 > x_1 x_2 x_3^3$ . Intuitively speaking, the Lex order views a polynomial  $f \in \mathbb{C}[\mathbf{x}]$  as a polynomial in  $x_1$  with coefficients that are polynomials in  $x_2$ , which has coefficients that are polynomials in  $x_3$ , and so on.  $\diamond$

Every monomial order induces the notion of *leading term* of a polynomial. Let  $f = \sum_{i=1}^k c_i \mathbf{x}^{\alpha_i} \in \mathbb{C}[\mathbf{x}]$ , where  $\alpha_1 > \alpha_2 > \dots > \alpha_k$  and  $c_1 \neq 0$ . Then, the leading term of  $f$  is  $\text{LT}(f) := c_1 \mathbf{x}^{\alpha_1}$ . The *leading term ideal* of an ideal  $I \subset \mathbb{C}[x_1, \dots, x_n]$  is defined as

$$\text{LT}(I) := \langle \{\text{LT}(f) \mid f \in I \setminus \{0\}\} \rangle.$$

Next comes the definition of a Gröbner basis.

**Definition 3.3 (Gröbner basis)** Let  $I \subset \mathbb{C}[x_1, \dots, x_n]$  be an ideal and  $>$  be a monomial order. A subset  $G = \{g_1, \dots, g_m\} \subset I$  is called a *Gröbner basis* for  $I$  with respect to  $>$  if its leading terms generate the leading term ideal; i.e., if

$$\langle \text{LT}(g_1), \dots, \text{LT}(g_m) \rangle = \text{LT}(I).$$

*Remark 3.4* If  $G$  is a Gröbner basis for an ideal  $I$ , then  $I = \langle G \rangle$  (see [54, Chapter 2 §5, Corollary 6]), hence the name ‘‘basis’’. A Gröbner basis defines a notion of *normal form* of an ideal. More specifically, let  $I \subset \mathbb{C}[\mathbf{x}]$  be an ideal and let  $G = \{g_1, \dots, g_m\}$  be a Gröbner basis for  $I$ . Then for every  $f \in \mathbb{C}[\mathbf{x}]$  there is a unique  $r \in \mathbb{C}[\mathbf{x}]$ , such that  $f = g + r$  with  $g \in I$  and no term of  $r$  is divisible by any of  $\text{LT}(g_1), \dots, \text{LT}(g_m)$  (see [54, Chapter 2 §6, Proposition 1]). The *remainder*  $r$  can be computed using the division algorithm.

Our next example is [54, Chapter 2 §8, Example 2]. It illustrates how Gröbner bases can be used to solve systems of polynomial equations.

**Example 3.5** Consider the following system of three polynomial equations in three variables:

$$F(x, y, z) = \begin{pmatrix} x^2 + y^2 + z^2 - 1 \\ x^2 + z^2 - y \\ x - z \end{pmatrix} = 0.$$

We compute a Gröbner basis of the ideal  $I = \langle x^2 + y^2 + z^2 - 1, x^2 + z^2 - y, x - z \rangle$  relative to the Lex order with  $x > y > z$  using Macaulay2 [79].

```
R = QQ[x, y, z, MonomialOrder => Lex];
f = x^2 + y^2 + z^2 - 1;
g = x^2 + z^2 - y;
h = x - z;
I = ideal {f, g, h};
G = gb I
gens G
```

This computes the Gröbner basis  $G = \{x - z, y - 2z^2, 4z^4 + 2z^2 - 1\}$ . Because  $I = \langle G \rangle$ , we can solve  $F(x, y, z) = 0$  by solving the system of equations given by  $G$ . Notice that the third polynomial in  $G$  only depends on  $z$ , the second only on  $y$  and  $z$ , and the third only on  $x$  and  $z$ . Thus, we can reduce solving  $F(x, y, z) = 0$  to solving three univariate polynomial equations. This gives us the four solutions  $(z, 2z^2, z)$ , where  $z$  iterates through the four roots of  $4z^4 + 2z^2 - 1$ .  $\diamond$

The reason why using the Lex order in Example 3.5 works well is the *Elimination Theorem*. To state this, let  $I \subset \mathbb{C}[\mathbf{x}]$  be an ideal. For every  $0 \leq j < n$ , the intersection  $I_j := I \cap \mathbb{C}[x_{j+1}, \dots, x_n]$  in an ideal in a polynomial subring. It consists of those polynomials in  $I$  that only contain the variables  $x_{j+1}, \dots, x_n$ . We call  $I_j$  the  $j$ -th *elimination ideal* of  $I$ . For a proof of the next theorem see [54, Chapter 3, §1, Theorem 1].

**Theorem 3.6 (The Elimination Theorem)** Let  $I \subset \mathbb{C}[\mathbf{x}]$  be an ideal and  $G$  be a Gröbner basis for  $I$  with respect to the Lex order with  $x_1 > \dots > x_n$ . Then

$$G_j := G \cap \mathbb{C}[x_{j+1}, \dots, x_n]$$

is a Gröbner basis of the  $j$ -th elimination ideal of  $I$ .

For us the most important consequence of the Elimination Theorem is that, if a system of polynomial equations  $F(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_n(\mathbf{x})) = 0$  has finitely many solutions, then the  $j$ -th elimination ideal will not be empty for  $0 \leq j < n$ . Consequently, we can solve  $F(\mathbf{x}) = 0$  by computing a Gröbner basis for the Lex order and then sequentially solving univariate equations. We can compute zeros of univariate polynomials by computing eigenvalues  $\lambda$  of the associated *companion matrix*. If  $f(x) = x^d + \sum_{i=0}^{d-1} c_i x^i$  is a univariate polynomial, we have  $f(\lambda) = 0$  if and only if  $\lambda$  is an eigenvalue of the companion matrix

$$\begin{pmatrix} 0 & \cdots & 0 & -c_0 \\ 1 & \cdots & 0 & -c_1 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & -c_{d-1} \end{pmatrix}.$$

Sometimes we are not interested in the solutions of  $F(\mathbf{x}) = 0$  per se, but only in the number of solutions. Gröbner bases naturally carry this information: Suppose  $I \subset \mathbb{C}[\mathbf{x}]$  is an ideal and  $>$  a monomial order. Recall that a monomial  $\mathbf{x}^\alpha \notin \text{LT}(I)$  is called a *standard monomial* of  $I$  relative to  $>$ . The next result shows how to get the number of solutions of  $F(\mathbf{x}) = 0$  from a Gröbner basis; see [170, Proposition 2.1].

**Proposition 3.7** *Let  $I \subset \mathbb{C}[\mathbf{x}]$  be an ideal. Let  $>$  be a term order and  $\mathcal{B}$  be the set of standard monomials of  $I$  relative to  $>$ . Then,  $\mathcal{B}$  is finite if and only if  $V(I)$  is finite, and  $\#\mathcal{B}$  equals the number of points in  $V(I)$  counting multiplicities.*

**Example 3.8** In Example 3.5, there are four standard monomials, namely, 1,  $z$ ,  $z^2$  and  $z^3$ . That is why the system  $F(x, y, z) = 0$  has four solutions.  $\diamond$

We have now understood how to solve systems of polynomials with finitely many zeros using Gröbner bases. What about systems whose variety has positive dimensional components? In that case, the  $n$ -th elimination ideal is necessarily the zero ideal. To cope with that case, one can remove positive dimensional components using ideal saturation. Let  $I, J \subset \mathbb{C}[\mathbf{x}]$  be two ideals. The *saturation* of  $I$  by  $J$  is the ideal

$$I : J^\infty := \{f \in \mathbb{C}[\mathbf{x}] \mid \text{there is } \ell > 0 \text{ with } f \cdot g^\ell \in I \text{ for all } g \in J\}. \quad (3.1)$$

Saturation is the ideal analogue of removing components on the level of varieties. We have

$$V(I : J^\infty) = \overline{V(I)} \setminus V(J); \quad (3.2)$$

see [54, Chapter 4 §4, Corollary 11].

Recall that a solution to a square system  $F(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_n(\mathbf{x})) = 0$  is called *regular* if the Jacobian determinant  $\det\left(\frac{\partial f_i}{\partial x_j}\right)_{1 \leq i, j \leq n}$  does not vanish at that solution. There can only be finitely many regular zeros of a square system of polynomial equations, and a finite union of points is Zariski closed. Consequently, if we saturate  $I = \langle f_1, \dots, f_n \rangle$  by  $J = \langle \det\left(\frac{\partial f_i}{\partial x_j}\right) \rangle$ , we can use the strategy from above to solve  $F(\mathbf{x}) = 0$ .

**Example 3.9** Consider the following system of two polynomials in two variables

$$F(x, y) = \begin{pmatrix} (x-1) \cdot (x-2) \cdot (x^2 + y^2 - 1) \\ (y-1) \cdot (y-3) \cdot (x^2 + y^2 - 1) \end{pmatrix} = 0.$$

It has 4 regular solutions  $(1, 1), (1, 3), (2, 1), (2, 3)$  and the circle  $x^2 + y^2 - 1$  as a positive dimensional component. We use Macaulay2 [79] to saturate the ideal generated by  $F$ .

```
R = QQ[x, y, MonomialOrder => Lex];
f = (x-1) * (x-2) * (x^2+y^2-1);
g = (y-1) * (y-3) * (x^2+y^2-1);
I = ideal {f, g};
Jac = matrix{{diff(x, f), diff(x, g)}, {diff(y, f), diff(y, g)}};
J = ideal det(Jac)
K = saturate(I, J)
```

This returns the ideal  $K = \langle y^2 - 4y + 3, x^2 - 3x + 2 \rangle$ . These two generators form a Gröbner basis for  $K$ .  $\diamond$

Next, we state two propositions related to elimination and saturation of ideals with parameters, and one lemma on Gröbner bases of parameterized ideals. For this, let  $\mathbf{x} = (x_1, \dots, x_n)$  and  $\mathbf{p} = (p_1, \dots, p_k)$  be two sets of variables, and  $\mathbb{C}[\mathbf{x}, \mathbf{p}] := \mathbb{C}[x_1, \dots, x_n, p_1, \dots, p_k]$ . We regard  $\mathbf{p}$  as variables for *parameters*. For a fixed parameter  $\mathbf{q} \in \mathbb{C}^k$ , we consider the surjective ring homomorphism

$$\phi_{\mathbf{q}} : \mathbb{C}[\mathbf{x}, \mathbf{p}] \rightarrow \mathbb{C}[\mathbf{x}], \quad f(\mathbf{x}; \mathbf{p}) \mapsto f(\mathbf{x}; \mathbf{q}). \quad (3.3)$$

**Proposition 3.10** Consider an ideal  $I \subset \mathbb{C}[\mathbf{x}, \mathbf{p}]$  and let  $G = \{g_1, \dots, g_m\}$  be a Gröbner basis for  $I$  relative to the Lex order  $x_1 > \dots > x_n > p_1 > \dots > p_k$ . For  $1 \leq i \leq m$  with  $g_i \notin \mathbb{C}[\mathbf{p}]$ , write  $g_i$  in the form  $g_i = c_i(\mathbf{p})\mathbf{x}^{\alpha_i} + h_i$ , where all terms of  $h_i$  are strictly smaller than  $\mathbf{x}^{\alpha_i}$ . Let  $\mathbf{q} \in V(I \cap \mathbb{C}[\mathbf{p}]) \subseteq \mathbb{C}^k$  such that  $c_i(\mathbf{q}) \neq 0$  for all  $g_i \notin \mathbb{C}[\mathbf{p}]$ . Then,

$$\phi_{\mathbf{q}}(G) = \{\phi_{\mathbf{q}}(g_i) \mid g_i \notin \mathbb{C}[\mathbf{p}]\}$$

is a Gröbner basis for the ideal  $\phi_{\mathbf{q}}(I) \subset \mathbb{C}[\mathbf{x}]$ .

**Proof** See, e.g., [54, Chapter 4 §7, Theorem 2].  $\square$

**Example 3.11 (Cardiod revisited)** As an illustration, fix  $n = m = 2$  and let  $I$  be the ideal generated by the cubic (2.4) and the determinant of (2.2), with  $u_1, u_2$  replaced by  $p_1, p_2$ . The lexicographic Gröbner basis  $G$  has 15 elements. The 15 leading coefficients  $c_i(\mathbf{p})$  are quite complicated. Three of them are

$$(4p_1^2 + 4p_2^2 + 4p_2 + 1)^2, \quad 32p_2^2(4p_2 + 3)(3p_2 + 1)(3p_2 + 2)^3(2p_2 + 1)^3(4p_2^2 + 5p_2 + 2)^3(1 + p_2)^5, \\ p_1p_2^2(4p_2 + 3)(8p_1^4 + 8p_1^2p_2^2 - 10p_1^2p_2 - 18p_2^3 - 7p_1^2 - 21p_2^2 - 8p_2 - 1)(p_1 + p_2 + 1)^3(p_1 - p_2 - 1)^3.$$

Suppose that we replace the unknowns  $p_1, p_2$  by any complex numbers  $q_1, q_2$  such that  $c_i(q_1, q_2) \neq 0$  for all  $i$ . Then the specialization  $\phi_{\mathbf{q}}(G)$  remains a Gröbner basis with the same leading terms. In particular, the number of zeros  $(x_1, x_2)$ , which is six when counted with multiplicities, is independent of  $q_1, q_2$ .

In many applications, polynomial systems come with a non-degeneracy constraint. These are usually expressed in the form of a polynomial inequation  $h \neq 0$ . We now show how to incorporate such a constraint.

**Proposition 3.12** Let  $I \subset \mathbb{C}[\mathbf{x}, \mathbf{p}]$  be an ideal and  $J = \langle h \rangle$  a principal ideal in the same ring. Let  $u$  be an additional variable and  $K := \langle 1 - u \cdot h \rangle$ . Then,

$$I : J^\infty = (I + K) \cap \mathbb{C}[\mathbf{x}, \mathbf{p}].$$

Furthermore, if  $G$  is a Gröbner basis of  $I + K$  relative to the Lex order  $u > x_1 > \dots > x_n > p_1 > \dots > p_k$ , then  $G \cap \mathbb{C}[\mathbf{x}, \mathbf{p}]$  is a Gröbner basis of  $I : J^\infty$ .

**Proof** See, e.g., [54, Chapter 4 §4, Theorem 14].  $\square$

**Example 3.13** Fix  $I$  from Example 3.11. Since the cardioid (2.4) is singular at the origin, we take  $K = \langle 1 - u \cdot h \rangle$  where  $h = x_1x_2$ . The Gröbner basis  $G$  has 27 elements, but with friendlier leading coefficients:

$$p_1p_2^5(p_1+p_2+1)(p_1-p_2-1), \quad p_2^5(p_2+1)^2, \quad p_1p_2^3, \quad p_2^2(9p_1^4+p_2^4+2p_2^3-9p_1^2+p_2^2), \quad p_2^2(p_2+1), \quad p_1(p_1^2-p_2-1), \dots$$

Nonvanishing of these coefficients ensures that the system has three complex solutions, with multiplicities.

We now apply the technique above to identifying a discriminant  $\Delta$ . The following is [24, Lemma 2.5].

**Lemma 3.14** Let  $I \subset \mathbb{C}[\mathbf{x}, \mathbf{p}]$  be an ideal and  $J = \langle h \rangle \subset \mathbb{C}[\mathbf{x}, \mathbf{p}]$  be a principal ideal such that  $(I : J^\infty) \cap \mathbb{C}[\mathbf{p}] = \{0\}$ . Let  $G = \{g_1, \dots, g_s\}$  be a Gröbner basis of  $I : J^\infty$  with respect to the Lex order  $x_1 > \dots > x_n > p_1 > \dots > p_k$ . There is a proper subvariety  $\Delta \subseteq \mathbb{C}^k$  such that, for all  $\mathbf{q} \notin \Delta$ , the set  $\{\phi_{\mathbf{q}}(g_1), \dots, \phi_{\mathbf{q}}(g_s)\}$  is a Gröbner basis for  $\phi_{\mathbf{q}}(I) : \phi_{\mathbf{q}}(J)^\infty$  and none of the leading terms of  $g_1, \dots, g_s$  vanish when evaluated at  $\mathbf{q}$ . In particular,  $\phi_{\mathbf{q}}(I : J^\infty) = \phi_{\mathbf{q}}(I) : \phi_{\mathbf{q}}(J)^\infty$  for all  $\mathbf{q} \notin \Delta$ .

**Proof** Let  $u$  be an additional variable and, as in Proposition 3.12, consider the ideal  $K := \langle 1 - u \cdot h \rangle$ . Then, we have  $I : J^\infty = (I + K) \cap \mathbb{C}[\mathbf{x}, \mathbf{p}]$ . Using our hypothesis  $(I : J^\infty) \cap \mathbb{C}[\mathbf{p}] = \{0\}$ , we conclude

$$V((I + K) \cap \mathbb{C}[\mathbf{p}]) = \mathbb{C}^k. \tag{3.4}$$

We may therefore apply Proposition 3.10 to  $I + K$  without any restrictions on  $\mathbf{q}$ . As in Proposition 3.12, we augment the Lex order by letting  $u$  be the largest variable. Let  $\overline{G} := \{g_1, \dots, g_r\}$  be a Gröbner basis of  $I + K$  relative to this order. By (3.4), we have  $g_1, \dots, g_r \notin \mathbb{C}[\mathbf{p}]$ . We write each  $g_i$  in the form  $g_i = c_i(\mathbf{p})u^\beta \mathbf{x}^{\alpha_i} + h_i$ , where all terms of  $h_i$  are strictly smaller than  $u^\beta \mathbf{x}^{\alpha_i}$ , and define the hypersurface

$$\Delta := \{\mathbf{q} \in \mathbb{C}^k \mid c_1(\mathbf{q}) \cdots c_r(\mathbf{q}) = 0\}. \quad (3.5)$$

Now let  $\mathbf{q} \in \mathbb{C}^k \setminus \Delta$ . By Proposition 3.10, the set  $\phi_{\mathbf{q}}(\overline{G}) = \{\phi_{\mathbf{q}}(g_1), \dots, \phi_{\mathbf{q}}(g_r)\}$  is a Gröbner basis for

$$\phi_{\mathbf{q}}(I + K) = \phi_{\mathbf{q}}(I) + \phi_{\mathbf{q}}(K) = \phi_{\mathbf{q}}(I) + (1 - u \cdot \phi_{\mathbf{q}}(h)).$$

Without loss of generality, suppose the first  $s \leq r$  elements in  $\overline{G}$  do not depend on the variable  $u$ . We define  $G := \{g_1, \dots, g_s\} = \overline{G} \cap \mathbb{C}[\mathbf{x}, \mathbf{p}]$ . It follows from Proposition 3.12 that  $G$  is a Gröbner basis of  $I : J^\infty$ . Because  $\mathbf{q} \notin \Delta$ , none of the leading terms in  $\overline{G}$  when evaluated at  $\mathbf{q}$  vanish. Consequently,

$$\phi_{\mathbf{q}}(G) \cap \mathbb{C}[\mathbf{x}] = \phi_{\mathbf{q}}(\overline{G}) \cap \mathbb{C}[\mathbf{x}].$$

Therefore,  $\phi_{\mathbf{q}}(G) = \{\phi_{\mathbf{q}}(g_1), \dots, \phi_{\mathbf{q}}(g_s)\}$  is a Gröbner basis of  $\phi_{\mathbf{q}}(I) : \phi_{\mathbf{q}}(J)^\infty$ , by Proposition 3.12.  $\square$

We refer to  $\Delta$  as the *discriminant* of the pair  $(I, h)$ . Equation (3.5) leads to the following corollary.

**Corollary 3.15** *Fix an ideal  $I = \langle f_1, \dots, f_n \rangle \subset \mathbb{C}[\mathbf{x}, \mathbf{p}]$  and a principal ideal  $J = \langle h \rangle \subset \mathbb{C}[\mathbf{x}, \mathbf{p}]$ . The discriminant  $\Delta$  in Lemma 3.14 is found by computing a Lex Gröbner basis  $G$  for  $I + \langle 1 - u \cdot h \rangle$ , where  $u > x_1 > \dots > x_n > p_1 > \dots > p_k$ . Namely,  $\Delta$  is the product of the leading coefficients  $c_i(\mathbf{p})$  in  $G$ .*

**Example 3.16 (Hyperdeterminant)** Let  $m = 8, n = 2$  and consider a general pair of bilinear equations:

$$I = \langle f_1, f_2 \rangle, \quad \text{where } f_1 = p_1x_1x_2 + p_2x_1 + p_3x_2 + p_4 \text{ and } f_2 = p_5x_1x_2 + p_6x_1 + p_7x_2 + p_8.$$

Our nondegeneracy constraint is the Jacobian determinant  $h = \partial f_1 / \partial x_1 \cdot \partial f_2 / \partial x_2 - \partial f_1 / \partial x_2 \cdot \partial f_2 / \partial x_1$ . The Gröbner basis  $G$  in Corollary 3.15 has 11 elements. The most interesting Gröbner basis element is

$$\begin{aligned} & c_1(\mathbf{p}) \cdot u - 2p_1p_7x_2 + 2p_3p_5x_2 - p_1p_8 - p_2p_7 + p_3p_6 + p_4p_5, \\ \text{where } & c_1(\mathbf{q}) = p_1^2p_8^2 - 2p_1p_2p_7p_8 - 2p_1p_3p_6p_8 - 2p_1p_4p_5p_8 + 4p_1p_4p_6p_7 + p_2^2p_7^2 \\ & + 4p_2p_3p_5p_8 - 2p_2p_3p_6p_7 - 2p_2p_4p_5p_7 + p_3^2p_6^2 - 2p_3p_4p_5p_6 + p_4^2p_5^2. \end{aligned}$$

This coefficient is the main factor in our discriminant  $\Delta$ . It is the *hyperdeterminant* of a  $2 \times 2 \times 2$  tensor.

## 3.2 The Parameter Continuation Theorem

The theorem to be proved in this section states that a square system of polynomial equations with parameters has a well-defined degree. This degree is the number of complex solutions for generic parameter choices. This result is the *Parameter Continuation Theorem* due to Morgan and Sommese [137]. We shall present a proof that rests on the results on Gröbner bases in the previous section. For this, we consider the polynomial ring  $\mathbb{C}[\mathbf{x}, \mathbf{p}] := \mathbb{C}[x_1, \dots, x_n, p_1, \dots, p_k]$ . We interpret  $\mathbf{x}$  as variables and  $\mathbf{p}$  as parameters.

**Definition 3.17** Let  $f_1(\mathbf{x}; \mathbf{p}), \dots, f_n(\mathbf{x}; \mathbf{p}) \in \mathbb{C}[\mathbf{x}, \mathbf{p}]$ . We consider the image of the polynomial map

$$\mathbb{C}^k \mapsto \mathbb{C}[\mathbf{x}]^n, \quad \mathbf{p}_0 \mapsto F(\mathbf{x}; \mathbf{p}_0) = \begin{pmatrix} f_1(\mathbf{x}; \mathbf{p}_0) \\ \vdots \\ f_n(\mathbf{x}; \mathbf{p}_0) \end{pmatrix}.$$

This image is a *family* of square polynomial systems of size  $n$ . In other words, the above family  $\mathcal{F} = \{F(\mathbf{x}; \mathbf{p}) \mid \mathbf{p} \in \mathbb{C}^k\}$  consists of  $n$  polynomials in  $n$  variables that depend polynomially on  $k$  parameters.

In Examples 3.11, 3.13 and 3.16, we studied families of polynomial systems with  $n = 2$ , where the number  $m$  of parameters was 2, 2 and 8, respectively. The degrees of these families are 6, 3 and 2. These degrees count the numbers of complex zeros for generic parameters  $\mathbf{p}$ , which satisfy  $c_i(\mathbf{p}) \neq 0$  for all  $i$ .

We now fix a family of polynomial systems  $\mathcal{F}$  depending on  $k$  parameters  $\mathbf{p} = (p_1, \dots, p_k)$ . Let  $\mathbf{z} \in \mathbb{C}^n$  be a zero of  $F(\mathbf{x}; \mathbf{p}) = (f_1(\mathbf{x}; \mathbf{p}), \dots, f_n(\mathbf{x}; \mathbf{p})) \in \mathcal{F}$ , for some specific parameters  $\mathbf{p} \in \mathbb{C}^k$ . We say that  $\mathbf{z}$  is a *regular zero* if the Jacobian determinant  $\det\left(\frac{\partial f_i}{\partial x_j}\right)_{1 \leq i, j \leq n}$  does not vanish at  $\mathbf{z}$ . The next theorem is our main result. It shows that, for almost all parameters  $\mathbf{p}$ , the number of regular solutions is the same.

**Theorem 3.18 (The Parameter Continuation Theorem)** *Let  $\mathcal{F}$  be a family of polynomial systems that consists of systems of  $n$  polynomials in  $n$  variables depending on  $k$  parameters. For  $\mathbf{p} \in \mathbb{C}^k$ , denote*

$$N(\mathbf{p}) := \#\{\mathbf{x} \in \mathbb{C}^n \mid \mathbf{x} \text{ is a regular zero of } F(\mathbf{x}; \mathbf{p})\}.$$

*Let  $N := \sup_{\mathbf{p} \in \mathbb{C}^k} N(\mathbf{p})$ . Then,  $N < \infty$ , and there exists a proper algebraic subvariety  $\Delta \subsetneq \mathbb{C}^k$ , called the discriminant of the system  $\mathcal{F}$ , such that  $N(\mathbf{p}) = N$  for  $\mathbf{p} \notin \Delta$ .*

**Proof** We recall the proof from [24]. Another proof can also be found in the textbook [166].

Suppose  $\mathcal{F} = \{F(\mathbf{x}; \mathbf{p}) \mid \mathbf{p} \in \mathbb{C}^k\}$ , where  $F(\mathbf{x}; \mathbf{p}) = (f_1(\mathbf{x}; \mathbf{p}), \dots, f_n(\mathbf{x}; \mathbf{p})) \in \mathcal{F}$ . Let

$$I = \langle f_1, \dots, f_n \rangle \quad \text{and} \quad J := \langle \det\left(\frac{\partial f_i}{\partial x_j}\right) \rangle.$$

If  $N = 0$ , then no system in  $\mathcal{F}$  has regular zeros. In this case, the statement is true. We now assume  $N > 0$ . By (3.2), the variety  $V(I : J^\infty)$  consists of all pairs  $(\mathbf{x}, \mathbf{q}) \in \mathbb{C}^n \times \mathbb{C}^k$  such that  $\mathbf{x}$  is a regular zero of  $F(\mathbf{x}; \mathbf{q})$ . Since  $N > 0$ , we therefore have  $V(I : J^\infty) \neq \emptyset$ . Let  $(\mathbf{x}, \mathbf{q}) \in V(I : J^\infty)$ . The Implicit Function Theorem ensures that there is a Euclidean open neighborhood  $U$  of  $\mathbf{q}$  such that  $F(\mathbf{x}; \mathbf{q})$  has regular zeros for all  $\mathbf{q} \in U$ . Consequently,  $(I : J^\infty) \cap \mathbb{C}[\mathbf{p}] = \{0\}$ , so we can apply Lemma 3.14 in our situation.

Set  $I_\mathbf{q} = \phi_\mathbf{q}(I)$  and  $J_\mathbf{q} = \phi_\mathbf{q}(J)$ . Let  $G = \{g_1, \dots, g_s\}$  be a Gröbner basis of  $I : J^\infty$  for the Lex order  $x_1 > \dots > x_n > p_1 > \dots > p_k$ . By Lemma 3.14, there is a proper algebraic subvariety  $\Delta \subsetneq \mathbb{C}^k$  such that  $\phi_\mathbf{q}(G) = \{\phi_\mathbf{q}(g_1), \dots, \phi_\mathbf{q}(g_s)\}$  is a Gröbner basis for  $I_\mathbf{q} : J_\mathbf{q}^\infty$  and none of the leading terms of  $g_1, \dots, g_s$  vanish when evaluated at  $\mathbf{q}$ . This implies that the leading monomials of  $I_\mathbf{q} : J_\mathbf{q}^\infty$  are constant on  $\mathbb{C}^k \setminus \Delta$ .

We consider the set of *standard monomials*, i.e. monomials not in the lexicographic initial ideal:

$$\mathcal{B}_\mathbf{q} := \{\text{standard monomials of } I_\mathbf{q} : J_\mathbf{q}^\infty\}.$$

We have shown is that  $\mathcal{B}_\mathbf{q}$  is constant on  $\mathbb{C}^k \setminus \Delta$ . On the other hand, by (3.2) and since finite sets of points are Zariski closed, we have  $V(I : J^\infty) = V(I) \setminus V(J)$ . Proposition 3.7 and the fact that regular zeros have multiplicity one imply that the following holds for all parameters  $\mathbf{q}$  that are not in the discriminant  $\Delta$ :

$$N(\mathbf{q}) = \#\mathcal{B}_\mathbf{q}.$$

This shows that  $N(\mathbf{q})$  is constant on  $\mathbb{C}^k \setminus \Delta$ . The Implicit Function Theorem implies that, for all  $\mathbf{q} \in \mathbb{C}^k$ , there exists a Euclidean neighborhood  $U$  of  $\mathbf{q}$  such that  $N(\mathbf{q}) \leq N(\mathbf{q}')$  for all  $\mathbf{q}' \in U$ . Since  $\Delta$  is a proper subvariety of  $\mathbb{C}^k$  and thus lower-dimensional, we have  $N = N(\mathbf{q}) < \infty$  for  $\mathbf{q} \in \mathbb{C}^k \setminus \Delta$ .  $\square$

We can use the algorithm in Corollary 3.15 to compute the discriminant. Observe that this algorithm also returns the discriminant when  $F(\mathbf{x}; \mathbf{p}) = 0$  has non-regular solutions for *all* parameters  $\mathbf{p}$ . Resultant-based methods for computing the discriminant would fail in such cases because the resultant will be constant and equal to zero. Here is a simple example to illustrate this phenomenon.

**Example 3.19** We slightly modify the system from Example 3.9 and consider

$$F(x, y; a) = \begin{pmatrix} (x-1) \cdot (x-2) \cdot (x^2+y^2-1) \\ (y-1) \cdot (y-a) \cdot (x^2+y^2-1) \end{pmatrix} = 0,$$

where  $a \in \mathbb{C}$  is a parameter. If  $a \notin \{0, 1, \pm\sqrt{-3}\}$ , we have  $N = 4$  regular solutions. Let us compute the discriminant using the algorithm in Corollary 3.15. We use the Macaulay2 code from [24].

```
R = QQ[u, x, y, a, MonomialOrder => Lex];
f = (x-1) * (x-2) * (x^2+y^2-1);
g = (y-1) * (y-a) * (x^2+y^2-1);
I = ideal {f, g};
Jac = matrix{{diff(x, f), diff(x, g)}, {diff(y, f), diff(y, g)}};
K = ideal {1 - u * det(Jac)};
G = gens gb (I+K);
E = (entries(G))#0

P = QQ[a][u, x, y, MonomialOrder => Lex]
result = apply(E, t -> leadCoefficient(sub(t, P)))
factor(product result)
```

The result is the polynomial  $a^6 \cdot (a-1)^4 \cdot (a^2+3)^2 \cdot g(a)$ , where  $g(a) = (a+1) \cdot (a^7 - 2a^6 + 8a^5 - 14a^4 + 23a^3 - 32a^2 + 32a - 32)$ . Each zero of the additional factor  $g(a)$  gives a system that also has four regular solutions. However, Theorem 3.18 only states that if a parameter is outside the discriminant, then it has the maximal number of regular zeros, but not the reverse implication.  $\diamond$

**Example 3.20** The critical equations for the ED problem in (2.1) have parameters  $\mathbf{u}$ . In many situations, the critical equations form a square system, and Theorem 3.18 applies. The degree  $N$  is the Euclidean Distance Degree. For instance, in Example 2.4, we have a square system in  $n = 3$  variables with  $m = 3$  parameters, and the ED degree equals  $N = d_1 d_2 (d_1 + d_2 + 1)$ . What is the discriminant  $\Delta$  in this case?  $\diamond$

**Example 3.21 (Tact invariant)** We consider the general system of two quadratic equations in  $n = 2$  variables. Each of the two equations has six coefficients, so there are  $m = 12$  parameters in total:

```
R = QQ[x, y, a20, a11, a02, a10, a01, a00, b20, b11, b02, b10, b01, b00];
f = a20*x^2 + a11*x*y + a02 * y^2 + a01*x + a10*y + a00;
g = b20*x^2 + b11*x*y + b02 * y^2 + b01*x + b10*y + b00;
```

For general parameter values  $\mathbf{q}$ , the set of lexicographic standard monomials is  $B_{\mathbf{q}} = \{1, y, y^2, y^3\}$ , so the number of solutions is  $N = 4$ . The main factor in the discriminant  $\Delta$  is a polynomial in the 12 coefficients that is known as the *tact invariant*. This polynomial can be computed with following Macaulay2 code:

```
I = ideal(f, g, diff(x, f)*diff(y, g) - diff(y, f)*diff(x, g));
tact = first first entries gens eliminate({x, y}, I);
toString tact
degree tact, # terms tact
```

From the output we see that the tact invariant has degree 12, and that it is the sum of 3210 monomials.

### 3.3 Polynomial Homotopy Continuation

In Section 3.1, we have seen how to use Gröbner bases to reduce solving  $F(\mathbf{x}) = 0$  to the problem of sequentially computing zeros of univariate polynomials. Another approach is *polynomial homotopy continuation* (PHC). This is a numerical method for computing the regular zeros of a square system of

polynomial equations. The textbook of Sommese and Wampler [166] provides a detailed introduction to the theory of polynomial homotopy continuation. We also refer to the overview article [14]. This subject area is known as *numerical algebraic geometry*, and the present section offers a lightning introduction.

The goal in polynomial homotopy continuation is to compute *approximate zeros*. The definition goes back to Smale (see [22, §8, Definition 1]). It rests on Newton's method from numerical analysis.

**Definition 3.22 (Approximate Zeros)** Let  $F(\mathbf{x})$  be a square system of polynomial equations in  $n$  variables, and write  $JF(\mathbf{x})$  for its (square) Jacobian matrix. A point  $\mathbf{z} \in \mathbb{C}^n$  is called an *approximate zero* of  $F$  if the sequence of Newton iterates  $\mathbf{z}_{k+1} := \mathbf{z}_k - JF(\mathbf{x}_k)^{-1}F(\mathbf{x}_k)$  starting at  $\mathbf{z}_0 := \mathbf{z}$  converges to a zero of  $F$ .

An approximate zero  $\mathbf{z}$  of a system  $F$  is in a precise sense close to an actual zero  $\mathbf{x}$ . Applying the Newton operator to  $\mathbf{z}$ , we can get as close to  $\mathbf{x}$  as we want. We can approximate  $\mathbf{x}$  to any desired accuracy.

Suppose now that  $F(\mathbf{x}) = 0$  is a system of polynomial equations that we want to solve. The idea in homotopy continuation is to find a family  $\mathcal{F}$  and parameters  $\mathbf{p}, \mathbf{q} \in \mathbb{C}^k$  with the following properties:

- $F(\mathbf{x}) = F(\mathbf{x}; \mathbf{p})$ ;
- $G(\mathbf{x}) := F(\mathbf{x}; \mathbf{q})$  is a system whose solutions are known or can be computed by other means.

For a (piecewise) smooth path  $\gamma(t)$  in  $\mathbb{C}^k$  with  $\gamma(0) = \mathbf{q}$  and  $\gamma(1) = \mathbf{p}$ , we define the *parameter homotopy*

$$H(\mathbf{x}, t) := F(\mathbf{x}; \gamma(t))$$

and *track* the solutions of  $F(\mathbf{x}; \mathbf{q}) = 0$  to  $F(\mathbf{x}; \mathbf{p})$  along the homotopy  $H$ . This tracking involves an ordinary differential equation (ODE). Namely, we use numerical algorithms to solve the *ODE initial value problem*

$$\left( \frac{d}{dx} H(\mathbf{x}, t) \right) \frac{dx}{dt} + \frac{d}{dt} H(\mathbf{x}, t) = 0, \quad \mathbf{x}(0) = \mathbf{z}. \quad (3.6)$$

Here the initial value  $\mathbf{z}$  is a zero of  $G(\mathbf{x})$ . In this setting,  $G(\mathbf{x}) = F(\mathbf{x}; \mathbf{q})$  is called *start system* and  $F(\mathbf{x}) = F(\mathbf{x}; \mathbf{p})$  is called *target system*. The output of the numerical solver is then an approximate zero of  $F(\mathbf{x})$ . In implementations one often uses piecewise linear paths, such as that described below.

**Remark 3.23** The left factor  $(\frac{d}{dx} H(\mathbf{x}, t))$  in (3.6) is the  $n \times n$  Jacobian matrix. Throughout the tracking process, it is essential that this matrix is invertible. This means geometrically that our path must stay away from the discriminant  $\Delta$  of the polynomial system. This is possible because  $\Delta$  is a proper subvariety of the parameter space  $\mathbb{C}^k$ . The dimension of  $\Delta$  over the real numbers  $\mathbb{R}$  is an even number that is less than the real dimension  $2k$  of the ambient space  $\mathbb{R}^{2k} = \mathbb{C}^k$ . This ensures that the space  $\mathbb{C}^k \setminus \Delta$  is connected.

The next proposition explains why polynomial homotopy continuation works and when the initial value problem from (3.6) is well-posed. The proof of the proposition crucially relies on the Parameter Continuation Theorem (Theorem 3.18), and on the connectedness result in Remark 3.23.

**Proposition 3.24** Let  $\mathcal{F}$  be a family of polynomial systems with parameters  $\mathbf{p} \in \mathbb{C}^k$ . Let  $N$  and  $\Delta$  be as in Theorem 3.18, and assume that  $N > 0$ . Given  $\mathbf{q} \in \mathbb{C}^k \setminus \Delta$  and  $\mathbf{p} \in \mathbb{C}^k$ , for almost all choices of  $\mathbf{p}_{\text{mid}} \in \mathbb{C}^k$ , the piecewise linear path

$$\gamma(t) = \begin{cases} (2t - 1)\mathbf{q} + 2(1 - t)\mathbf{p}_{\text{mid}}, & \text{if } \frac{1}{2} \leq t \leq 1 \\ 2t\mathbf{p}_{\text{mid}} + (1 - 2t)\mathbf{p}, & \text{if } 0 < t \leq \frac{1}{2} \end{cases}$$

satisfies:

1.  $\gamma((0, 1]) \cap \Delta = \emptyset$ .

2. The homotopy  $H(\mathbf{x}, t) := F(\mathbf{x}; \gamma(t))$  defines  $N$  smooth curves  $\mathbf{x}(t)$  with the property that  $H(\mathbf{x}(t), t) = 0$  for  $0 < t \leq 1$ . These curves are called solution paths.
3. As  $t \rightarrow 0$ , the limits of the solution paths include all regular solutions of  $F(\mathbf{x}; \mathbf{p}) = 0$ .
4. If moreover  $\gamma(0) \notin \Delta$ , then every solution path  $\mathbf{x}(t)$  converges for  $t \rightarrow 0$  to a regular zero of  $F(\mathbf{x}; \mathbf{p})$ .

**Proof** If  $N > 0$ , the discriminant  $\Delta \subseteq \mathbb{C}^k$  is a proper complex subvariety. Therefore,  $\Delta$  is of complex codimension at least 1, hence of real codimension at least 2. This was a recap of Remark 3.23. It implies that for general  $\mathbf{p}_{\text{mid}} \in \mathbb{C}^k$ , the path  $\gamma(t)$  does not intersect  $\Delta$  for  $t \in (0, 1]$ . This proves the first item.

Since  $\mathbf{q} = \gamma(0) \notin \Delta$ , the implicit function theorem ensures that there exists a Euclidean neighborhood  $\mathcal{U}_0 \subset \mathbb{C}^k$  of  $\mathbf{q}$  and a smooth solution map  $s_0 : \mathcal{U}_0 \rightarrow \mathbb{C}^n$  such that  $F(s_0(\mathbf{p}), \mathbf{p}) = 0$  for all  $\mathbf{p} \in \mathcal{U}_0$ . Let

$$t_0 := \min \{t \in [0, 1] \mid \gamma(t) \in \overline{\mathcal{U}_0}\},$$

where  $\overline{\mathcal{U}_0}$  is the Euclidean closure of  $\mathcal{U}_0$ . If  $t_0 > 0$ , then  $\gamma(t_0) \notin \Delta$  and we can repeat the construction for the new start system  $F(\mathbf{x}; \gamma(t_0))$ . Eventually, we obtain an open cover

$$(0, 1] = \bigcup_{i \in \mathcal{I}} \mathcal{U}_i$$

for some index set  $\mathcal{I}$ , together with smooth solution maps  $s_i : \mathcal{U}_i \rightarrow \mathbb{C}^n$ . Taking a partition of unity  $(\rho_i(t))_{i \in \mathcal{I}}$  relative to this cover (see [124, Chapter 2]), we set

$$\mathbf{x}(t) := \sum_{i \in \mathcal{I}} \rho_i(t) \cdot (s_i \circ \gamma)(t), \quad t \in (0, 1].$$

By construction, the path  $\mathbf{x}(t)$  is smooth and has the property that  $H(\mathbf{x}(t), t) = 0$ . Furthermore, as  $t \rightarrow 0$ , the solution path  $\mathbf{x}(t)$  either converges to a point  $\mathbf{z}$  or diverges as  $\|\mathbf{x}_i(t)\| \rightarrow \infty$ . In the first situation, by continuity,  $\mathbf{z}$  is a zero of  $F(\mathbf{x}, \gamma(0)) = F(\mathbf{x}, \mathbf{p})$  (not necessarily regular).

By Theorem 3.18,  $F(\mathbf{x}; \mathbf{q})$  has  $N$  regular zeros. The construction above yields  $N$  solution paths  $\mathbf{x}_1(t), \dots, \mathbf{x}_N(t)$ . Smoothness implies that  $\mathbf{x}_i(t) \neq \mathbf{x}_j(t)$  for  $i \neq j$  and all  $t \in (0, 1]$ . This proves the second item. Furthermore, for every regular zero of  $F(\mathbf{x}; \mathbf{q})$ , we also find a (local) solution map, which connects to exactly one of the smooth paths  $\mathbf{x}_1(t), \dots, \mathbf{x}_N(t)$ . This implies the third and fourth item.  $\square$

There are several software packages for solving polynomial systems that are based on homotopy continuation. In this book we use the software `HomotopyContinuation.jl` [29] that was developed by the first author together with Sascha Timme. In our view, this software is quite powerful and easy to use.

**Example 3.25** We use to solve the system of polynomial equations from Example 3.5. The following are commands in the programming language `Julia` on which `HomotopyContinuation.jl` is based:

```
using HomotopyContinuation
@var x y z;
f = x^2 + y^2 + z^2 - 1;
g = x^2 + z^2 - y;
h = x - z;
F = System([f; g; h], variables = [x; y; z])
solve(F)
```

This code returns the four solutions (here displayed with only the 4 most significant digits):

$$\begin{aligned}
& (0.556 + 0.0\sqrt{-1}, 0.618 - 0.0\sqrt{-1}, 0.556 + 0.0\sqrt{-1}), \\
& (-0.0 - 0.899\sqrt{-1}, -1.618 + 0.0\sqrt{-1}, -0.0 - 0.899\sqrt{-1}), \\
& (-0.556 - 0.0\sqrt{-1}, 0.618 + 0.0\sqrt{-1}, -0.556 + 0.0\sqrt{-1}), \\
& (0.0 + 0.899\sqrt{-1}, -1.618 + 0.0\sqrt{-1}, -0.0 + 0.899\sqrt{-1}).
\end{aligned}$$

These are numerical approximations of the solutions found symbolically in Example 3.5.  $\diamond$

*Remark 3.26* The capabilities of `HomotopyContinuation.jl` were already on display in Example 2.8. In that example we solved the critical equations for the ED problem on some complete intersections.

*Remark 3.27* Numerical computations are not exact computations and therefore can produce errors. This is hence also true for polynomial homotopy continuation. It is possible, though, to *certify* the output of PHC. Certification means that we obtain a computer proof that we have indeed computed an approximate zero. There are various certification methods. Current implementations are [28, 87, 125].

**Corollary 3.28** *A general system of polynomials  $F(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_n(\mathbf{x}))$  in  $n$  variables has*

$$N = d_1 \cdots d_n$$

*isolated zeros in  $\mathbb{C}^n$ , where  $d_i = \deg f_i$ . (The number  $d_1 \cdots d_n$  is also called the Bézout number of  $F$ .)*

**Proof** We consider the family  $\mathcal{F}_{\text{Bézout}}$  of polynomial systems  $F(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_n(\mathbf{x}))$  with  $d_i = \deg f_i$ . The parameters are the coefficients of the polynomials  $f_1, \dots, f_n$ . Here, we can use the start system

$$G(\mathbf{x}) = \begin{pmatrix} x_1^{d_1} - 1 \\ \vdots \\ x_n^{d_n} - 1 \end{pmatrix}.$$

This system has the  $d_1 \cdots d_n$  distinct complex zeros. Explicitly, the zeros are  $(\xi_1^{k_1}, \dots, \xi_n^{k_n})$ , where  $\xi_i := \exp(2\pi\sqrt{-1}/d_i)$  is the  $d_i$ -th root of unity and  $k_i$  ranges from 1 to  $d_i$ . One calls  $G(x)$  the *total degree start system*. All its zeros are regular, and it has no zeros at infinity. Together with Proposition 3.24 this implies that the system  $G(\mathbf{x})$  has the maximal number  $N = d_1 \cdots d_n$  of regular zeros in  $\mathcal{F}_{\text{Bézout}}$ .  $\square$

*Remark 3.29* Corollary 3.28 only states that the number of solutions equals  $d_1 \cdots d_n$  for systems outside the discriminant. The full version of Bézout's theorem also applies to systems  $F \in \Delta$ , where it states that the number of zeros counted with multiplicities is  $d_1 \cdots d_n$ . Corollary 3.28 does not prove this full version.

Corollary 3.28 implies that one can use the total degree start system for homotopy continuation in the family of system of polynomials with fixed degree pattern.

**Example 3.30** The system  $F(x, y, z) = (x^2 + y^2 + z^2 - 1, x^2 + z^2 - y, x - z)$  from Example 3.25 consists of three polynomials of degrees  $d_1 = 2, d_2 = 2$  and  $d_3 = 1$ . We have shown that the number of zeros of  $F$  is the Bézout number  $N = d_1 \cdot d_2 \cdot d_3 = 4$ . In `HomotopyContinuation.jl` [29], we can use the total degree start system by setting the following flag:

```
solve(F; start_system = :total_degree)
```

The default option in `HomotopyContinuation.jl` is the *polyhedral homotopy*. Here the start system is constructed from Newton polytopes, and it respects the mixed volume, as described from Example 3.31.  $\diamond$

**Example 3.31** Let  $A \subset \mathbb{N}^n$  be a finite set and denote  $\mathcal{F}_A := \{\sum_{\alpha \in A} c_\alpha \mathbf{x}^\alpha \mid c_\alpha \in \mathbb{C}\}$ . An element in  $\mathcal{F}_A$  is called a *sparse polynomial*, since only the monomials with exponent vector in  $A$  appear. For finite subsets  $A_1, \dots, A_n \subset \mathbb{N}^n$ , we consider the family  $\mathcal{F}_{\text{sparse}} := \mathcal{F}_{A_1} \times \dots \times \mathcal{F}_{A_n}$ . The parameters in this family are the coefficients of the  $n$  sparse polynomials of a system in  $\mathcal{F}_{\text{sparse}}$ . For  $1 \leq i \leq n$ , let  $P_i$  be the convex hull of  $A_i$ . The polytope  $P_i$  is called the *Newton polytope* of the polynomial  $F_{A_i}$ .

Let  $\text{MV}(P_1, \dots, P_n)$  denote the *mixed volume* of these  $n$  polytopes. The BKK Theorem [19, 20] asserts that a general  $F \in \mathcal{F}_{\text{sparse}}$  has  $\text{MV}(P_1, \dots, P_n)$  many zeros in the torus  $(\mathbb{C}^*)^n$ . Assuming that a general  $F \in \mathcal{F}_{\text{sparse}}$  only has zeros with nonzero entries, the maximal number of regular zeros in  $\mathcal{F}_{\text{sparse}}$  therefore is

$$N = \text{MV}(P_1, \dots, P_n).$$

If the supports  $A_i$  are all the same, then there is only one Newton polytope  $P = P_1 = \dots = P_n$ . In this situation, which occurs frequently in practise, the mixed volume  $\text{MV}(P, \dots, P)$  equals  $n!$  times the volume of  $P$ . This product is a nonnegative integer, and it is referred to as the normalized volume of  $P$ . Thus the BKK bound for unmixed square systems is the normalized volume of the Newton polytope.  $\diamond$

**Remark 3.32** The article [103] provides an algorithm to compute an explicit start system for  $\mathcal{F}_{\text{sparse}}$ , called *polyhedral start system*. See also the summary in [14, Section 3]. The use of the polyhedral start system is the default option in `HomotopyContinuation.jl`.

We close this chapter with a brief discussion for two quadratic equations in two variables. The general system appeared in Example 3.21, where we computed the tact invariant, which serves as the discriminant. The nonvanishing of the tact invariant ensures that the two equations have four distinct complex solutions. Suppose we begin with the total degree start system  $x_1^2 = x_2^2 = 1$ , which has four solutions  $(\pm 1, \pm 1)$ . The homotopy in Proposition 3.24 is guaranteed to find the four solutions of the system we wish to solve.

By contrast, suppose now that our two quadratic equations are sparse, and the system has the form

$$F(x, y) = \begin{pmatrix} a + bx + cy + dxy \\ \alpha + \beta x + \gamma y + \delta xy \end{pmatrix} \in \mathcal{F}_{\text{sparse}},$$

where  $\mathbf{p} = (a, b, c, d, \alpha, \beta, \gamma, \delta) \in \mathbb{C}^8$  are parameters. Both polynomials in  $F$  have the same Newton polytope  $P$ , namely the unit square. The normalized volume of the unit square equals  $\text{MV}(P) = 2$ . Therefore, the BKK Theorem tells us that  $F(x, y) = 0$  has  $N = 2$  solutions for general parameters  $\mathbf{p} \in \mathbb{C}^8$ .

The total degree start system  $x_1^2 = x_2^2 = 1$  is not appropriate for the sparse family  $F(x, y)$  because it has too many solutions. Indeed, two of the start solutions  $(\pm 1, \pm 1)$  lead to paths that diverge when running the homotopy. Instead, one can use a polyhedral start system [103] that has precisely two solutions. A polyhedral start system is obtained by dividing the square  $P$  into two triangles, each of normalized area 1.

The two zeros of  $F(x, y)$  are distinct when the discriminant does not vanish at  $\mathbf{p}$ . The discriminant of the system  $F(x, y)$  is the hyperdeterminant of format  $2 \times 2 \times 2$ , which we computed in Example 3.16. In other words, the role of the tact invariant for two dense quadrics is now played by our hyperdeterminant. Hyperdeterminants of larger tensors, and the spectral theory of tensors, will be discussed in later chapters.

## **Chapter 4**

### **Polar Degrees**

We compare three definitions of *polar degrees*, in terms of non-transversal intersections, Schubert varieties and the Gauss map, and conormal varieties. The latter approach was used in Chapter 2.3. We discuss the key properties of polar degrees under projective duality and explain how polar degrees are related with Chern classes.

We work over an algebraically closed field of characteristic zero.

## 4.1 Polar Varieties

**Example 4.1** Imagine that you look at an algebraic surface  $X \subseteq \mathbb{P}^3$  from a point  $V \in \mathbb{P}^3$ . If you would want to sketch the surface from your point of view, you would draw its *contour curve*  $P(X, V)$ ; see Figure 4.1. The contour curve consists of all points  $p$  on the surface  $X$  such that the line spanned by  $V$  and  $p$  is tangent at  $p$ . The *first polar degree*  $\mu_1(X)$  of the surface  $X$  is the degree of the contour curve for a generic point  $V$ .

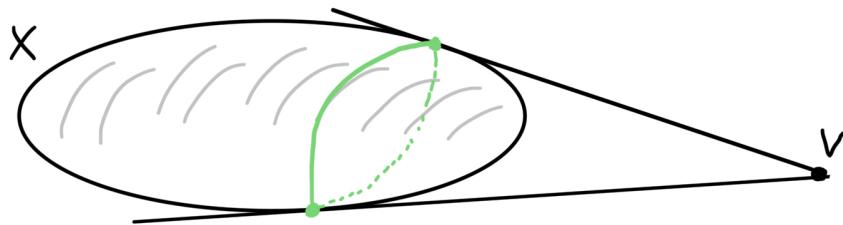


Fig. 4.1: Green contour curve when the ellipsoid  $X$  is viewed from the point  $V$ .

Now we change the setting slightly and imaging that our viewing of the surface  $X$  is not centered at a point but at a line  $V \subseteq \mathbb{P}^3$ . This time our contour set  $P(X, V)$  consists of all point  $p$  on the surface  $X$  such that the plane spanned by the point  $p$  and the line  $V$  is tangent at  $p$ ; see Figure 4.2. For a generic line  $V$ , the contour set  $P(X, V)$  is finite, and the *second polar degree* is its cardinality.  $\diamond$

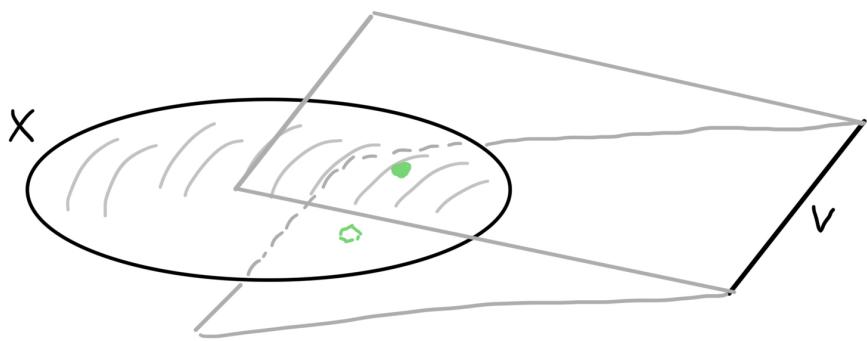


Fig. 4.2: The green contour set consists of two points when the ellipsoid  $X$  is viewed from the line  $V$ .

The contour sets described in the example above are also known as *polar varieties*. To define polar varieties in general, we need to fix some conventions and notations. For instance, the dimension of the empty set is considered to be  $-1$ . Given projective subspaces  $V, W \subseteq \mathbb{P}^n$ , their projective span (equivalently, their join) is denoted by  $V + W \subseteq \mathbb{P}^n$ . If  $V$  and  $W$  are disjoint in projective space then we have

$$\dim(V + W) = \dim(V) + \dim(W) + 1$$

Given a projective variety  $X \subseteq \mathbb{P}^n$ , we write  $\text{Reg}(X)$  for its regular locus. The *embedded tangent space* of  $X$  at  $p \in \text{Reg}(X)$  is

$$\mathbb{T}_p X := \left\{ v \in \mathbb{P}^n \mid \forall f \in I(X) : \sum_{i=0}^n \frac{\partial f}{\partial x_i}(p) \cdot v_i = 0 \right\}.$$

We recall that a projective subspace  $W \subseteq \mathbb{P}^n$  is said to intersect  $X$  *non-transversely* at  $p \in \text{Reg}(X)$  if  $p \in W$  and  $\dim(W + \mathbb{T}_p X) < n$ . For instance, if  $X$  is a smooth curve in  $\mathbb{P}^3$ , then every line that intersects it does so non-transversely, while the tangent planes of  $X$  are the only planes that meet  $X$  non-transversely.

**Definition 4.2** Let  $X \subseteq \mathbb{P}^n$  be an irreducible projective variety. The *polar variety* of  $X$  with respect to a projective subspace  $V \subseteq \mathbb{P}^n$  is

$$P(X, V) := \overline{\{p \in \text{Reg}(X) \setminus V \mid V + p \text{ intersects } X \text{ at } p \text{ non-transversely}\}}.$$

For every  $i \in \{0, \dots, \dim X\}$ , there is an integer  $\mu_i(X)$  that is equal to the degree of  $P(X, V)$  for almost all projective subspaces  $V \subseteq \mathbb{P}^n$  with  $\dim V = \text{codim } X - 2 + i$ . The nonnegative integer  $\mu_i(X)$  is called the *i-th polar degree* of  $X$ .

**Example 4.3** A surface  $X \subseteq \mathbb{P}^3$  has three polar degrees. For  $i = 0, 1, 2$ , the generic subspace  $V$  in Definition 4.2 is empty, a point, or a line, respectively. We saw the latter two cases in Example 4.1. For the case  $i = 0$ , we observe that  $P(X, \emptyset) = X$ . So, the 0-th polar degree  $\mu_0(X)$  is the degree of the surface  $X$ .  $\diamond$

**Example 4.4** The last observation that  $\mu_0(X) = \deg(X)$  is true in general. If  $i = 0$ , the dimension of the generic subspace  $V$  in Definition 4.2 is  $\text{codim } X - 2$ . Hence, we have for every  $p \in \text{Reg}(X)$  that

$$\begin{aligned} \dim((V + p) + \mathbb{T}_p X) &= \dim(V + \mathbb{T}_p X) = \dim V + \dim X - \dim(V \cap \mathbb{T}_p X) \\ &\leq (\text{codim } X - 2) + \dim X + 1 = n - 1, \end{aligned}$$

which means that  $V + p$  intersects  $X$  at  $p$  non-transversely. Therefore, we conclude that  $P(X, V) = X$  and  $\mu_0(X) = \deg(X)$ .  $\diamond$

We will now give a second definition of polar varieties in terms of the Gauss map and Schubert varieties. For that fix a projective subspace  $V \subseteq \mathbb{P}^n$ . We observe that  $V + p$  intersects  $X$  at  $p \in \text{Reg}(X)$  non-transversely (i.e.,  $n > \dim((V + p) + \mathbb{T}_p X) = \dim V + \dim X - \dim(V \cap \mathbb{T}_p X)$ ) if and only if

$$\dim(V \cap \mathbb{T}_p X) > \dim V - \text{codim } X. \tag{4.1}$$

Since  $\dim V - \text{codim } X$  is the expected dimension of the intersection of the two projective subspaces  $V$  and  $\mathbb{T}_p X$ , condition (4.1) means that the tangent space  $\mathbb{T}_p X$  meets  $V$  in an unexpectedly large dimension. Such subspaces are collected in simple instances of *Schubert varieties*:

$$\Sigma_m(V) := \{T \in \text{Gr}(m, \mathbb{P}^n) \mid \dim(V \cap T) > \dim V - n + m\}.$$

If  $m = \dim X$ , then condition (4.1) is equivalent to  $\mathbb{T}_p X \in \Sigma_m(V)$ . Hence, the *Gauss map*

$$\begin{aligned}\gamma_X : X &\dashrightarrow \mathrm{Gr}(m, \mathbb{P}^n), \\ p &\longmapsto \mathbb{T}_p X\end{aligned}$$

pulls the Schubert variety  $\Sigma_m(V)$  back to the polar variety  $P(X, V)$ , i.e.,

$$P(X, V) = \overline{\gamma_X^{-1}(\Sigma_{\dim X}(V))}.$$

## 4.2 Projective Duality

We recall that there is a one-to-one correspondence between hyperplanes  $H$  in the  $\mathbb{P}^n$  and points  $H^\vee$  in the dual projective space  $(\mathbb{P}^n)^*$ . The *dual variety* of a projective variety  $X \subseteq \mathbb{P}^n$  consists of all tangent hyperplanes of  $X$ :

$$X^\vee := \overline{\{H^\vee \in (\mathbb{P}^n)^* \mid \exists p \in \mathrm{Reg}(X) : \mathbb{T}_p X \subseteq H\}}.$$

**Theorem 4.5 (Biduality theorem, [77])** Let  $X \subseteq \mathbb{P}^n$  be a projective variety over an algebraically closed field of characteristic zero. Moreover, let  $p \in \mathrm{Reg}(X)$  and  $H^\vee \in \mathrm{Reg}(X^\vee)$ . The hyperplane  $H$  is tangent to  $X$  at the point  $p$  if and only if the hyperplane  $p^\vee$  is tangent to  $X^\vee$  at the point  $H^\vee$ . In particular,  $(X^\vee)^\vee = X$ .

**Example 4.6** If  $X$  is a projective subspace of  $\mathbb{P}^n$ , then  $X^\vee$  is a projective subspace of  $(\mathbb{P}^n)^*$  with  $\dim X^\vee = n - 1 - \dim X$ . In particular, we have that  $(\mathbb{P}^n)^\vee = \emptyset$ .

**Example 4.7** Let us revisit the example in Figure 4.2 that illustrates the polar variety  $P(X, V)$  of a surface  $X \subseteq \mathbb{P}^3$  and a generic line  $V$ . The tangent planes passing through the line  $V$  correspond in  $(\mathbb{P}^3)^*$  to points on the dual variety  $X^\vee$  that are contained in the line  $V^\vee$ ; see Figure 4.3. Hence, if the dual variety is a surface as well, then its degree is given by the second polar degree of  $X$ , i.e.,  $\mu_2(X) = \deg(X^\vee)$ . Otherwise, if the dual variety  $X^\vee$  is of smaller dimension, the line  $V^\vee$  misses it and  $\mu_2(X) = 0$ .

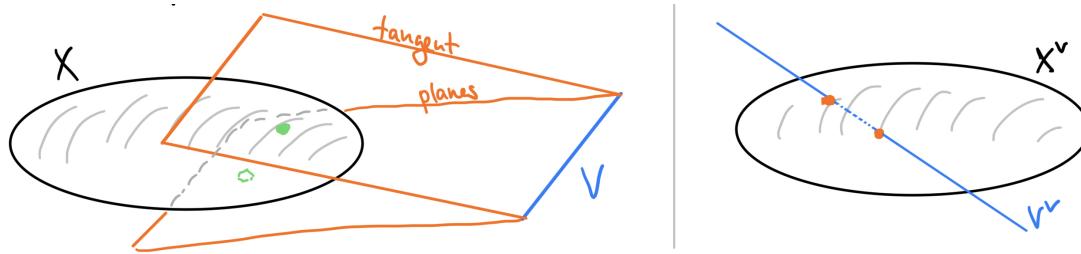


Fig. 4.3:  $\mu_2(X) = \deg(X^\vee)$  holds for pairs of dual surfaces in projective 3-space ( $q_i = (\mathbb{T}_p X)^\vee$ ).

In the setting of Figure 4.1, the polar variety  $P(X, V)$  of a surface  $X \subseteq \mathbb{P}^3$  and a generic point  $V$  is a curve. Thus, its degree ( $= \mu_1(X)$ ) is computed by intersecting it with a generic plane  $H$ . The polar curve consists of all points on  $X$  whose tangent plane contains the point  $V$ , i.e.,  $P(X, V) = \overline{\{p \in \mathrm{Reg}(X) \mid V \in \mathbb{T}_p X\}}$ . Hence, the first polar degree  $\mu_1(X)$  counts all (regular) points  $p \in X$  such that

$$p \in H \text{ and } V \in \mathbb{T}_p X. \quad (4.2)$$

The tangent planes at those points correspond to points  $q := (\mathbb{T}_p X)^\vee$  in the dual projective space. By the biduality theorem, those points satisfy  $\mathbb{T}_q X^\vee = p^\vee$  if the dual variety  $X^\vee$  is a surface. Hence, in that case, the two conditions in (4.2) are equivalent to

$$H^\vee \in \mathbb{T}_q X^\vee \text{ and } q \in V^\vee. \quad (4.3)$$

Comparing now (4.3) with (4.2), we see that the point-plane pair  $(H^\vee, V^\vee)$  imposes the same conditions on the points  $q \in X^\vee$  as the point-plane pair  $(V, H)$  imposes on the points  $p \in X$ ; see also Figure 4.4. Due to the genericity of  $(V, H)$ , we conclude that  $\mu_1(X) = \mu_1(X^\vee)$  if  $X^\vee$  is a surface. If  $X^\vee$  is a curve, then  $\mathbb{T}_q X^\vee \subseteq p^\vee$  and so the only conditions imposed by (4.2) on the points  $q \in X^\vee$  is that they must lie in  $V^\vee$ , i.e.,  $\mu_1(X) = |X^\vee \cap V^\vee| = \deg(X^\vee) = \mu_0(X^\vee)$ . Finally, if  $X^\vee$  is a point (i.e.,  $X$  is a plane), then  $\mu_1(X) = 0$ .  $\diamond$

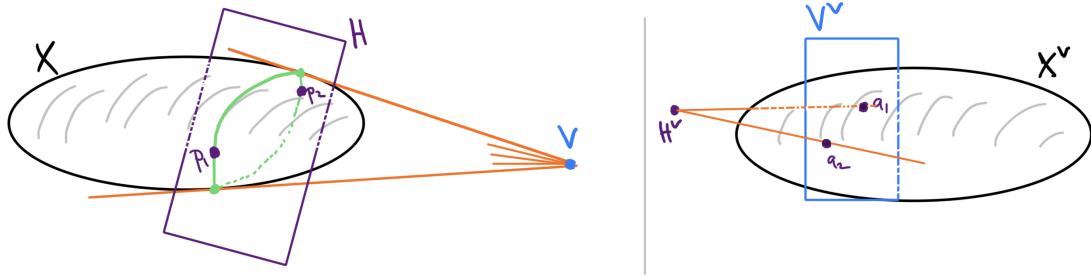


Fig. 4.4:  $\mu_1(X) = \mu_1(X^\vee)$  holds for pairs of dual surfaces in projective 3-space.

The relations between the polar degrees of a variety and its dual that we observed in the previous example are true in more generality. More specifically, the list of polar degrees of a projective variety  $X$  is exactly the list of polar degrees of its dual variety  $X^\vee$  in reversed order. As seen in Example 4.4, the first non-zero entry in that list is  $\deg(X)$ , and so its last non-zero entry is  $\deg(X^\vee)$ . We summarize these key properties of polar degrees:

**Proposition 4.8 ([95])** *Let  $X$  be an irreducible projective variety, and let  $\alpha(X) := \dim X - \operatorname{codim} X^\vee + 1$ .*

- (a)  $\mu_i(X) > 0 \iff 0 \leq i \leq \alpha(X)$ .
- (b)  $\mu_0(X) = \deg X$ .
- (c)  $\mu_{\alpha(X)}(X) = \deg X^\vee$ .
- (d)  $\mu_i(X) = \mu_{\alpha(X)-i}(X^\vee)$ .

The ideas discussed in Example 4.7 can be turned into formal proofs for almost all assertions in Proposition 4.8 (only the direction “ $\Leftarrow$ ” in (a) is rather tricky). Another strategy is to first establish the relation of the polar degrees with the *conormal variety* of the projective variety  $X \subseteq \mathbb{P}^n$ :

$$\mathcal{N}_X := \overline{\{(p, H^\vee) \in \mathbb{P}^n \times (\mathbb{P}^n)^* \mid p \in \operatorname{Reg}(X), \mathbb{T}_p X \subseteq H\}}.$$

The projection of the conormal variety  $\mathcal{N}_X$  onto the first resp. second factor is the variety  $X$  resp. its dual  $X^\vee$ . Moreover, by the biduality theorem, we have that

$$\mathcal{N}_{X^\vee} = \{(H^\vee, p) \mid (p, H^\vee) \in \mathcal{N}_X\}. \quad (4.4)$$

Independently of the dimension of  $X$ , the dimension of the conormal variety is always  $n - 1$ . Hence, the *multidegree* of the conormal variety is given by its intersections with  $L_1 \times L_2$ , where  $L_1 \subseteq \mathbb{P}^n$  and  $L_2 \subseteq (\mathbb{P}^n)^*$  are generic subspaces with  $\dim L_1 + \dim L_2 = \text{codim } \mathcal{N}_X = n + 1$ . We denote the entries of that multidegree by

$$\delta_j(X) := |\mathcal{N}_X \cap (L_1 \times L_2)|, \text{ for generic } L_1, L_2 \text{ with } \dim L_2 = j, \dim L_1 = n + 1 - j.$$

We saw in Section 2.3 that the multidegree is the cohomology class of the conormal variety  $\mathcal{N}_X$ .

**Proposition 4.9** ([111, Prop. (3) on page 187] or [75, Lem. (2.23) on page 169]) *The multidegree agrees with the polar degrees. More precisely, we have  $\delta_j(X) = \mu_i(X)$ , where  $i := \dim X + 1 - j$ .*

This proposition together with (4.4) implies immediately Proposition 4.8(d) and hence using Example 4.4 also (b) and (c). The direction “ $\Rightarrow$ ” in Proposition 4.8(a) can also be deduced directly from the definition of the  $\delta_j(X)$ .

Before we present an idea of proof for Proposition 4.9, we revisit our running example.

**Example 4.10** We see from Figure 4.4 and the conditions (4.2) and (4.3) that the first polar degree of a surface  $X$  in  $\mathbb{P}^3$  is computed as  $\mu_1(X) = |\mathcal{N}_X \cap (H \times V^\vee)| = \delta_2(X)$ . In other words, Proposition 4.8 holds for surfaces  $X$  in  $\mathbb{P}^3$ .

**Proof (Sketch for Proposition 4.9)** Let  $L_1 \subseteq \mathbb{P}^n$  and  $L_2 \subseteq (\mathbb{P}^n)^*$  be generic subspaces of dimensions  $n + 1 - j$  and  $j$ , respectively. Setting  $V := L_2^\vee$ , we start by observing that  $V$  has the correct dimension to be used in the computation of the  $i$ -th polar degree (where  $i = \dim X + 1 - j$ ), since  $\dim V = n - j - 1 = \text{codim } X - 2 + i$ .

Now we consider a generic pair  $(p, H^\vee) \in \mathcal{N}_X \cap (\mathbb{P}^n \times L_2)$ . The point  $p \in X$  is regular and both its tangent space  $\mathbb{T}_p X$  and  $V = L_2^\vee$  are contained in the hyperplane  $H$ . In particular, we have  $\dim(V + \mathbb{T}_p X) < n$  and so  $p$  is in the polar variety  $P(X, V)$ . In fact, the projection  $\mathcal{N}_X \cap (\mathbb{P}^n \times L_2) \rightarrow P(X, V)$  onto the first factor is birational. Hence,  $\mu_i(X) = \deg P(X, V) = |P(X, V) \cap L_1| = |\mathcal{N}_X \cap (L_1 \times L_2)| = \delta_j(X)$ .  $\square$

### 4.3 Chern Classes

Chern classes are topological invariants associated with vector bundles on smooth manifolds or varieties. For a smooth, irreducible projective variety  $X$ , its polar degrees can be computed from its Chern classes (see Proposition 4.11).

To a vector bundle  $\mathcal{E}$  on  $X$  of rank  $r$ , we associate the Chern classes  $c_0(\mathcal{E}), \dots, c_r(\mathcal{E})$ , which are formally elements in the *Chow ring* of  $X$ . Chern classes are easiest understood when the vector bundle  $\mathcal{E}$  is globally generated. In that case, the Chern class  $c_{r+1-j}(\mathcal{E})$  is the element in the Chow ring of  $X$  that is associated with the following degeneracy locus:

$$D(\sigma_1, \dots, \sigma_j) := \{x \in X \mid \sigma_1(x), \dots, \sigma_j(x) \text{ are linearly dependent}\},$$

where  $\sigma_1, \dots, \sigma_j : X \rightarrow \mathcal{E}$  are  $j$  general global sections. For the purpose of this section, it is not crucial to understand the Chow ring. It suffices for us to understand the *degree* of  $c_{r+1-j}(\mathcal{E})$ . This is defined to be the degree of the degeneracy locus  $D(\sigma_1, \dots, \sigma_j)$  for general  $\sigma_i$ . For instance, the degree of the *top Chern class*  $c_r(\mathcal{E})$  is the degree of the vanishing locus of a single general global section.

There are some calculation rules that allow us to compute Chern classes of more complex vector bundles. Most notably, the *Whitney sum formula* states for a short exact sequence  $0 \rightarrow \mathcal{E}' \rightarrow \mathcal{E} \rightarrow \mathcal{E}'' \rightarrow 0$  of

vector bundles that  $c_k(\mathcal{E}) = \sum_{i+j=k} c_i(\mathcal{E}')c_j(\mathcal{E}'')$ ; see [74, Theorem 3.2]. The Chern class  $c_k(X)$  of  $X$  is an abbreviation for the Chern class  $c_k(\mathcal{T}X)$  of its tangent bundle  $\mathcal{T}X$ .

**Proposition 4.11** ([95, eq. (3)]) *Let  $X$  be a smooth, irreducible projective variety, and let  $m := \dim X$ . Then,*

$$\mu_i(X) = \sum_{k=0}^i (-1)^k \binom{m-k+1}{m-i+1} \deg(c_k(X)).$$

This formula can also be reverted to express degrees of Chern classes in terms of polar degrees:

$$\deg(c_k(X)) = \sum_{i=0}^k (-1)^i \binom{m-i+1}{m-k+1} \mu_i(X). \quad (4.5)$$

*Remark 4.12* Both formulas also hold for singular varieties, after replacing the classical Chern classes with Chern-Mather classes. That result is due to R. Piene (see [150, Theorem 3] or [149]).

An important difference between polar degrees and Chern classes is the following: Polar degrees are projective invariants of the embedded variety  $X \subseteq \mathbb{P}^n$ . This holds also more generally for the *polar classes*, i.e., the rational equivalence classes (in the Chow ring of  $X$ ) of the polar varieties. Chern classes are even *intrinsic invariants* of the variety  $X$ , i.e., they do not depend on the embedding of  $X$  in projective space.

**Example 4.13** Let  $X$  be a smooth, irreducible projective variety.

- a) We see from (4.5) that  $\deg(c_0(X)) = \mu_0(X) = \deg X$ .
- b) The top Chern class of  $X$  coincides with its topological Euler characteristic:  $\deg(c_m(X)) = \chi(X)$ , where  $m = \dim X$ .
- c) If  $X$  is a curve, then  $\chi(X) = 2 - 2g(X)$ , where  $g(X)$  denotes the genus. Moreover, we see from (4.5) that  $\deg(c_1(X)) = 2\deg X - \mu_1(X)$ . Hence, we conclude:

$$2\deg X - \mu_1(X) = 2 - 2g(X). \quad (4.6)$$

- d) If  $X \subseteq \mathbb{P}^n$  is a rational curve, we can easily verify the relation (4.6):

- If  $X$  is a line, its dual variety is never a hypersurface, and so  $\mu_1(X) = 0$ .
- If  $X$  is a conic (i.e.,  $\deg X = 2$ ), its dual variety is (a cone over) a conic, and so  $\mu_1(X) = \deg X^\vee = 2$ .
- If  $X$  is a twisted cubic (i.e.,  $\deg X = 3$ ), its dual variety is (a cone over) the discriminant hypersurface of a cubic polynomial, and so  $\mu_1(X) = \deg X^\vee$  is the degree of that discriminant, which is 4.
- More generally, if  $X$  is a rational normal curve of degree  $d$ , its dual variety is (a cone over) the discriminant hypersurface of a degree- $d$  polynomial, and so  $\mu_1(X) = \deg X^\vee$  is the degree of that discriminant, which is  $2d - 2$ .



## **Chapter 5**

### **Wasserstein Distance**

A fundamental problem in metric algebraic geometry is distance minimization. We seek a point in a variety  $X$  in  $\mathbb{R}^n$  that is closest to a given data point  $\mathbf{u} \in \mathbb{R}^n$ . Thus, we must solve the optimization problem

$$\text{minimize } \|\mathbf{x} - \mathbf{u}\| \text{ subject to } \mathbf{x} \in X. \quad (5.1)$$

In what follows, this minimum in (5.1) is always attained because  $X$  is non-empty and closed. Hence there exists at least one optimal solution. If that solution is unique then we denote it by  $\mathbf{x}^*$ .

In the previous chapter we discussed this problem for the Euclidean norm on  $\mathbb{R}^n$ . In what follows we study (5.1) in the case when the distance is given by a *polyhedral norm*. We will then focus on a particular class of polyhedral norms that arise from optimal transport theory. These are known as Wasserstein norms.

## 5.1 Polyhedral Norms

A norm  $\|\cdot\|$  on the real vector space  $\mathbb{R}^n$  is a *polyhedral norm* if its unit ball is polyhedral:

$$B = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| \leq 1\}.$$

More precisely,  $B$  is a centrally symmetric convex polytope. Conversely, every centrally symmetric convex polytope  $B$  in  $\mathbb{R}^n$  defines a polyhedral norm on  $\mathbb{R}^n$ . Using the unit ball, we can paraphrase (5.1) as follows:

$$\text{minimize } \lambda \text{ subject to } \lambda \geq 0 \text{ and } (\mathbf{u} + \lambda B) \cap X \neq \emptyset. \quad (5.2)$$

Familiar examples of polyhedral norms are  $\|\cdot\|_\infty$  and  $\|\cdot\|_1$ . For these norms, the unit ball  $B$  is, respectively, the cube and the crosspolytope (a.k.a. generalized octahedron). Polyhedral norms are very important in optimal transport theory, where one uses a Wasserstein norm on the space of probability distributions. This will be our main application in this chapter, and it will be discussed in detail in the later sections.

We begin our discussion with a general polyhedral norm, that is, we allow  $B$  to be an arbitrary  $n$ -dimensional centrally symmetric polytope in  $\mathbb{R}^n$ . The boundary of  $B$  consists of faces whose dimensions range from 0 to  $n - 1$ . We use the dot  $\cdot$  for the standard inner product on  $\mathbb{R}^n$ . Recall that a subset  $F$  of the polytope  $B$  is a *face* if there exists a linear functional  $\ell \in \mathbb{R}^n \setminus \{0\}$  such that

$$F = \{\mathbf{x} \in B : \ell \cdot \mathbf{x} \geq \ell \cdot \mathbf{y} \text{ for all } \mathbf{y} \in B\}. \quad (5.3)$$

The set of all faces, ordered by inclusion, is a partial ordered set, called the *face poset* of  $B$ . An important combinatorial invariant of our polytope  $B$  is its *f-vector*  $f(B) = (f_0, f_1, \dots, f_{n-1})$ . By definition, the  $i$ th coordinate  $f_i$  of the f-vector is the number of  $i$ -dimensional faces of  $B$ .

The dual of the unit ball  $B$  is also a centrally symmetric polytope, namely it is the set

$$B^* = \{\ell \in \mathbb{R}^n : \ell \cdot \mathbf{x} \leq 1 \text{ for all } \mathbf{x} \in B\}.$$

The norm  $\|\cdot\|_*$  defined by the dual polytope  $B^*$  is dual to the norm  $\|\cdot\|$  given by  $B$ . The f-vector of  $B^*$  is the reverse of the f-vector of  $B$ . More precisely, we have  $f_i(B^*) = f_{n-1-i}(B)$  for  $i = 0, 1, \dots, n - 1$ .

**Example 5.1** Fix the unit cube  $B = [-1, 1]^n$ . Its dual is the crosspolytope

$$B^* = \text{conv}\{\pm \mathbf{e}_1, \pm \mathbf{e}_2, \dots, \pm \mathbf{e}_n\} \subset \mathbb{R}^n.$$

Here  $\mathbf{e}_j$  is the  $j$ th standard basis vector. The number of  $i$ -dimensional faces of the cube is

$$f_i(B) = \binom{n}{i} \cdot 2^{n-i}.$$

The 3-dimensional crosspolytope is the octahedron. The 3-cube and the octahedron satisfy

$$f(B) = (8, 12, 6) \quad \text{and} \quad f(B^*) = (6, 12, 8).$$

These numbers govern the combinatorial structure of the norms  $\|\cdot\|_\infty$  and  $\|\cdot\|_1$  on  $\mathbb{R}^3$ .

We now turn to the critical equations for the optimization problem given in (5.1) or (5.2). To derive these equations, a combinatorial stratification of the problem will be used. This is given by the face poset of the polytope  $B$ . Suppose that the variety  $X$  is in sufficiently general position in  $\mathbb{R}^n$ . This hypothesis implies that  $(\mathbf{u} + \lambda^* B) \cap X = \{\mathbf{x}^*\}$  is a singleton for the optimal value  $\lambda^*$  in (5.2). The point  $\frac{1}{\lambda^*}(\mathbf{x}^* - \mathbf{u})$  lies in boundary of the unit ball  $B$ . Hence it lies in the relative interior of a unique face  $F$  of the polytope  $B$ . Let  $L_F$  denote the linear span of  $F$  in  $\mathbb{R}^n$ . We have  $\dim(L_F) = \dim(F) + 1$ . Let  $\ell$  be any linear functional on  $\mathbb{R}^n$  that attains its maximum over the polytope  $B$  at the face  $F$ . This means that (5.3) holds.

**Lemma 5.2** *The optimal point  $\mathbf{x}^*$  in (5.1) is the unique solution to the optimization problem*

$$\text{Minimize } \ell(\mathbf{x}) \text{ subject to } \mathbf{x} \in (\mathbf{u} + L_F) \cap X. \quad (5.4)$$

**Proof** The general position hypothesis ensures that the affine space  $\mathbf{u} + L_F$  intersects the real variety  $X$  transversally, and  $\mathbf{x}^*$  is a smooth point of that intersection. Moreover,  $\mathbf{x}^*$  is a minimum of the restriction of  $\ell$  to the variety  $(\mathbf{u} + L_F) \cap X$ . By our hypothesis, this linear function is generic relative to the variety, so the number of critical points is finite and the function values are distinct.  $\square$

**Example 5.3 (Touching at a facet)** Suppose that the face  $F$  is a facet of the unit ball  $B$ . Then  $L_F = \mathbb{R}^n$ , and  $\ell$  is an outer normal vector to that facet, which is unique up to scaling. Here, the optimization problem (5.4) asks for the minimum of  $\ell$  over  $X$ . This situation corresponds to the left diagram in Figure 5.1.

**Example 5.4 (Touching at a vertex)** Suppose  $F$  is a vertex of the unit ball  $B$ . This case arises when  $X$  is a hypersurface. It corresponds to the middle diagram in Figure 5.1. Here, the affine space  $\mathbf{u} + L_F$  is the line that connects  $\mathbf{u}$  and  $\mathbf{x}^*$ . That line intersects  $X$  in a finite set of cardinality  $\text{degree}(X)$ . The optimal  $\mathbf{x}^*$  is the real point in that finite set at which the value of the linear form  $\ell$  is minimal.

Problem (5.4) amounts to linear programming over a real variety. We now determine the algebraic degree of this optimization task when  $F$  is a face of codimension  $i$ . To this end, we replace the affine variety  $X \subset \mathbb{R}^n$  by its closure in complex projective space  $\mathbb{P}^n$ . We retain the same symbol  $X$  for that projective variety. Consider the affine space  $L = u + L_F$  in  $\mathbb{R}^n$ , and also identify it with its closure in  $\mathbb{P}^n$ . If the face  $F$  has codimension  $i$  then the linear space  $L$  has codimension  $i - 1$ . The following result assumes that this space is in general position relative to the variety  $X$  and relative to the isotropic quadric.

**Theorem 5.5** *Let  $L$  be a general affine-linear space of codimension  $i - 1$  in  $\mathbb{R}^n$  and let  $\ell$  be a general linear form. The number of critical points of  $\ell$  on  $L \cap X$  is the polar degree  $\delta_i(X)$ .*

**Proof** This result appears in [46, Theorem 5.1]. The number of critical points of a linear form is the degree of the dual variety  $(L \cap X)^\vee$ . That degree coincides with the polar degree  $\delta_i(X)$ .  $\square$

**Example 5.6** Examples 5.3 and 5.4 explain Theorem 5.5 in the two extreme cases  $i = 1$  and  $i = n$ . Touching at a vertex ( $i = n$ ) can only happen when  $X$  is a hypersurface, and here  $\delta_n(X) = \text{degree}(X)$ . Touching at a facet ( $i = 1$ ) can happen for varieties of any dimension, as long as the dual variety  $X^\vee$  is a hypersurface. In that case we have  $\delta_1(X) = \text{degree}(X^\vee)$ .

Theorem 5.5 offers a direct interpretation of each polar degree  $\delta_i(X)$  in terms of optimization on  $X$ . This interpretation can be used as a definition of polar degrees. Some readers might prefer this.

**Example 5.7** Consider the distance minimization problem in (5.1) and (5.2) where  $X$  is a general surface of degree  $d$  in  $\mathbb{R}^3$ . The optimal face  $F$  of the unit ball  $B$  depends on the location of the data point  $\mathbf{u}$ . The algebraic degree of the solution  $\mathbf{x}^*$  equals  $\delta_3(X) = d$  if  $\dim(F) = 0$ , it is  $\delta_2(X) = d(d - 1)$  if  $\dim(F) = 1$ , and it is  $\delta_1(X) = d(d - 1)^2$  if  $\dim(F) = 2$ . Here  $\mathbf{x}^*$  is the unique point in  $(\mathbf{u} + \lambda^* B) \cap X$ , where  $\lambda^*$  is the optimal value in (5.2). Figure 5.1 visualizes this scenario for  $d = 2$  and  $\|\cdot\|_\infty$ . The variety  $X$  is the green sphere, which is a surface of degree  $d = 2$ . The unit ball for the norm  $\|\cdot\|_\infty$  is the cube  $B = [-1, 1]^3$ . The picture shows the smallest  $\lambda^*$  such that  $\mathbf{u} + \lambda^* B$  touches the sphere  $X$ . The cross marks the point of contact. This is the point  $\mathbf{x}^*$  in  $X$  which is closest in  $\infty$ -norm to the green point  $\mathbf{u}$  in the center of the cube. Point of contact is either on a facet, or on an edge, or it is a vertex. The algebraic degree of  $\mathbf{x}^*$  is two in all three cases, i.e. we can write the solution  $\mathbf{x}^*$  in terms of the data  $\mathbf{u}$  by solving the quadratic formula. If the green surface in Figure 5.1 were a cubic surface then these polar degrees would be 3, 6 and 12.

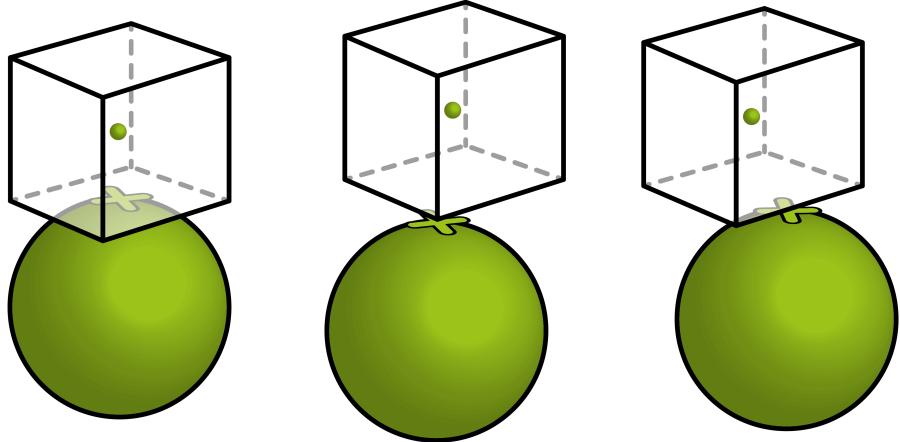


Fig. 5.1: The cube is an  $\|\cdot\|_\infty$  ball around the green point  $u$ . The variety  $X$  is the sphere. The contact point  $x^*$  is marked with a cross. The optimal face  $F$  is a facet, vertex, or edge.

We have learned that the conormal variety  $N_X$  and its cohomology class  $[N_X]$  are key players when it comes to reliably solving the distance minimization problem for a variety  $X$ . This applies not just to the Euclidean distance problem, but also to the analogous problem for polyhedral norms. The polar degrees  $\delta_i(X)$  reveal precisely how many paths need to be tracked by numerical solvers like [15, 34] in order to find and certify [33] the optimal solution  $\mathbf{x}^*$  in (5.1) or (5.4).

## 5.2 Optimal Transport and Independence Models

We now come to the main theme of this chapter, namely the Wasserstein distance to a given variety  $X$ . For us,  $X$  will be an independence model in a probability simplex, described algebraically by matrices or tensors of low rank, and we measure distances using Wasserstein metrics on that simplex. This is a class of polyhedral norms which are important in optimal transport theory. We now present the relevant definitions.

A probability distribution on the finite set  $[n] = \{1, 2, \dots, n\}$  is a point  $\nu$  in the simplex  $\Delta_{n-1} = \{(\nu_1, \dots, \nu_n) \in \mathbb{R}_{\geq 0}^n : \sum_{i=1}^n \nu_i = 1\}$ . We metrize this simplex by the *Wasserstein distance*. To define this

notion, we first turn the state space  $[n]$  into a finite metric space by fixing a symmetric  $n \times n$  matrix  $d = (d_{ij})$  with nonnegative entries. These entries satisfy  $d_{ii} = 0$  and  $d_{ik} \leq d_{ij} + d_{jk}$  for all  $i, j, k$ . Given two probability distributions  $\mu$  and  $\nu$  in  $\Delta_{n-1}$ , we consider the following linear programming problem, where  $\mathbf{z} = (z_1, \dots, z_n)$  denotes the decision variables:

$$\text{Maximize } \sum_{i=1}^n (\mu_i - \nu_i) z_i \text{ subject to } |z_i - z_j| \leq d_{ij} \text{ for all } 1 \leq i < j \leq n. \quad (5.5)$$

The optimal value of (5.5), denoted  $W_d(\mu, \nu)$ , is the *Wasserstein distance* between  $\mu$  and  $\nu$ .

The optimal solution  $\mathbf{z}^*$  to problem (5.5) is known as the *optimal discriminator* for the two probability distributions  $\mu$  and  $\nu$ . It satisfies  $W_d(\mu, \nu) = \langle \mu - \nu, \mathbf{z}^* \rangle$ , and its coordinates  $z_i^*$  are weights on the state space  $[n]$  that tell  $\mu$  and  $\nu$  apart. Here  $\langle \cdot, \cdot \rangle$  is the standard inner product on  $\mathbb{R}^n$ . The linear program (5.5) is the *Kantorovich dual* of the *optimal transport problem*.

The feasible region of the linear program (5.5) is unbounded because it is invariant under translation by  $\mathbf{1} = (1, 1, \dots, 1)$ . It is compact after taking the quotient modulo the line  $\mathbb{R}\mathbf{1}$ :

$$P_d = \{ \mathbf{z} \in \mathbb{R}^n / \mathbb{R}\mathbf{1} : |z_i - z_j| \leq d_{ij} \text{ for all } 1 \leq i < j \leq n \}. \quad (5.6)$$

This  $(n-1)$ -dimensional polytope is the *Lipschitz polytope* of the metric space  $([n], d)$ . In the field of tropical geometry, one calls  $P_d$  a *polytrope* because it is convex both classically and tropically.

The polytope  $P_d^*$  that is dual to  $P_d$  lies in the hyperplane perpendicular to the line  $\mathbb{R}\mathbf{1}$ . We call  $P_d^*$  the *root polytope* because its vertices are, up to scaling, the elements  $\mathbf{e}_i - \mathbf{e}_j$  in the root system of Lie type  $A_{n-1}$ . More precisely, we have

$$P_d^* = \{ x \in \mathbb{R}^n : \max_{z \in P_d} \langle x, z \rangle \leq 1 \} = \text{conv} \left\{ \frac{1}{d_{ij}} (\mathbf{e}_i - \mathbf{e}_j) : 1 \leq i, j \leq n \right\}.$$

This is a centrally symmetric polytope since the finite metric space  $([n], d)$  satisfies  $d_{ij} = d_{ji}$ .

**Proposition 5.8** *The Wasserstein metric  $W_d$  on the probability simplex  $\Delta_{n-1}$  is given by the polyhedral norm whose unit ball is the root polytope  $P_d^*$ .*

**Proof** Fix the polyhedral norm with unit ball  $P_d^*$ . The distance between  $\mu$  and  $\nu$  in this norm is the smallest real number  $\lambda$  such that  $\mu \in \nu + \lambda P_d^*$ , or, equivalently,  $\frac{1}{\lambda}(\mu - \nu) \in P_d^*$ . By definition of dual polytope, this minimal  $\lambda$  is the maximum inner product  $\langle \mu - \nu, z \rangle$  over all points  $z$  in the dual  $(P_d^*)^*$  of the unit ball. But this specifies the Lipschitz polytope, i.e.  $(P_d^*)^* = P_d$ . Hence the distance between  $\mu$  and  $\nu$  is equal to  $W_d(\mu, \nu)$ , which is the optimal value in (5.5).  $\square$

**Example 5.9** Let  $n = 4$  and fix the finite metric space graph distance on the 4-cycle

$$d = \begin{pmatrix} 0 & 1 & 1 & 2 \\ 1 & 0 & 2 & 1 \\ 1 & 2 & 0 & 1 \\ 2 & 1 & 1 & 0 \end{pmatrix}. \quad (5.7)$$

The induced metric on the tetrahedron  $\Delta_3$  is given by the Lipschitz polytope

$$\begin{aligned} P_d &= \{ (x_1, x_2, x_3, x_4) \in \mathbb{R}^4 / \mathbb{R}\mathbf{1} : |x_1 - x_2| \leq 1, |x_1 - x_3| \leq 1, |x_2 - x_4| \leq 1, |x_3 - x_4| \leq 1 \} \\ &= \text{conv} \{ (1, 0, 0, -1), (-1, 0, 0, 1), (\frac{1}{2}, -\frac{1}{2}, -\frac{1}{2}, \frac{1}{2}), (-\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, -\frac{1}{2}), (0, 1, -1, 0), (0, -1, 1, 0) \}. \end{aligned}$$

Note that this 3-dimensional polytope is an octahedron. Therefore, its dual is a cube:

$$\begin{aligned} P_d^* &= \{(y_1, y_2, y_3, y_4) \in (\mathbb{R}1)^\perp : |y_1 - y_4| \leq 1, |y_2 - y_3| \leq 1, |y_2 + y_3| \leq 1\} \\ &= \text{conv}\{(1, -1, 0, 0), (1, 0, -1, 0), (0, 1, 0, -1), (0, 0, 1, -1) \\ &\quad (-1, 1, 0, 0), (-1, 0, 1, 0), (0, -1, 0, 1), (0, 0, -1, 1)\}. \end{aligned}$$

This is the unit ball for the Wasserstein metric on the tetrahedron  $\Delta_3$  that is induced by  $d$ . Measuring the distance from a point to a surface with respect to this metric is illustrated in Figure 5.1.

We wish to compute the Wasserstein distance from a given distribution  $\mu$  to a fixed *discrete statistical model*  $\mathcal{M} \subset \Delta_{n-1}$ . This is the problem studied in [45, 46]. Our discussion serves as an introduction. As is customary in algebraic statistics, we assume that  $\mathcal{M}$  is defined by polynomials in  $v_1, \dots, v_n$ .

Our task is to solve the following optimization problem:

$$W_d(\mu, \mathcal{M}) := \min_{v \in \mathcal{M}} W_d(\mu, v) = \min_{v \in \mathcal{M}} \max_{x \in P_d} \langle \mu - v, x \rangle. \quad (5.8)$$

Computing this quantity means solving a non-convex optimization problem. Our aim is to study this problem and propose solution strategies, using methods from geometry, algebra and combinatorics. The analogous problem for the Euclidean metric was treated earlier.

We now present a detailed case study for the tetrahedron  $\Delta_3$  whose points are joint probability distributions of two binary random variables. The *2-bit independence model*  $\mathcal{M} \subset \Delta_3$  consists of all nonnegative  $2 \times 2$  matrices of rank one whose entries sum to one:

$$\begin{pmatrix} v_1 & v_2 \\ v_3 & v_4 \end{pmatrix} = \begin{pmatrix} pq & p(1-q) \\ (1-p)q & (1-p)(1-q) \end{pmatrix}, \quad (p, q) \in [0, 1]^2. \quad (5.9)$$

Thus,  $\mathcal{M}$  is the quadratic surface in the tetrahedron  $\Delta_3$  defined by the equation  $v_1 v_4 = v_2 v_3$ . The next theorem gives the optimal value function and the solution function for this independence model. We use the Wasserstein metric  $W_d$  that was defined in Example 5.9.

**Theorem 5.10** *The Wasserstein distance from a distribution  $\mu \in \Delta_3$  to the surface  $\mathcal{M}$  equals*

$$W_d(\mu, \mathcal{M}) = \begin{cases} 2\sqrt{\mu_1}(1 - \sqrt{\mu_1}) - \mu_2 - \mu_3 & \text{if } \mu_1 \geq \mu_4, \sqrt{\mu_1} \geq \mu_1 + \mu_2, \sqrt{\mu_1} \geq \mu_1 + \mu_3, \\ 2\sqrt{\mu_2}(1 - \sqrt{\mu_2}) - \mu_1 - \mu_4 & \text{if } \mu_2 \geq \mu_3, \sqrt{\mu_2} \geq \mu_1 + \mu_2, \sqrt{\mu_2} \geq \mu_2 + \mu_4, \\ 2\sqrt{\mu_3}(1 - \sqrt{\mu_3}) - \mu_1 - \mu_4 & \text{if } \mu_3 \geq \mu_2, \sqrt{\mu_3} \geq \mu_1 + \mu_3, \sqrt{\mu_3} \geq \mu_3 + \mu_4, \\ 2\sqrt{\mu_4}(1 - \sqrt{\mu_4}) - \mu_2 - \mu_3 & \text{if } \mu_4 \geq \mu_1, \sqrt{\mu_4} \geq \mu_2 + \mu_4, \sqrt{\mu_4} \geq \mu_3 + \mu_4, \\ |\mu_1\mu_4 - \mu_2\mu_3|/(\mu_1 + \mu_2) & \text{if } \mu_1 \geq \mu_4, \mu_2 \geq \mu_3, \mu_1 + \mu_2 \geq \sqrt{\mu_1}, \mu_1 + \mu_2 \geq \sqrt{\mu_2}, \\ |\mu_1\mu_4 - \mu_2\mu_3|/(\mu_1 + \mu_3) & \text{if } \mu_1 \geq \mu_4, \mu_3 \geq \mu_2, \mu_1 + \mu_3 \geq \sqrt{\mu_1}, \mu_1 + \mu_3 \geq \sqrt{\mu_3}, \\ |\mu_1\mu_4 - \mu_2\mu_3|/(\mu_2 + \mu_4) & \text{if } \mu_4 \geq \mu_1, \mu_2 \geq \mu_3, \mu_2 + \mu_4 \geq \sqrt{\mu_4}, \mu_2 + \mu_4 \geq \sqrt{\mu_2}, \\ |\mu_1\mu_4 - \mu_2\mu_3|/(\mu_3 + \mu_4) & \text{if } \mu_4 \geq \mu_1, \mu_3 \geq \mu_2, \mu_3 + \mu_4 \geq \sqrt{\mu_4}, \mu_3 + \mu_4 \geq \sqrt{\mu_3}. \end{cases}$$

The solution function  $\Delta_3 \rightarrow \mathcal{M}$ ,  $\mu \mapsto v^*(\mu)$  is given (with the same case distinction) by

$$v^*(\mu) = \begin{cases} (\mu_1, \sqrt{\mu_1} - \mu_1, \sqrt{\mu_1} - \mu_1, -2\sqrt{\mu_1} + \mu_1 + 1), \\ (\sqrt{\mu_2} - \mu_2, \mu_2, -2\sqrt{\mu_2} + \mu_2 + 1, \sqrt{\mu_2} - \mu_2), \\ (\sqrt{\mu_3} - \mu_3, -2\sqrt{\mu_3} + \mu_3 + 1, \mu_3, \sqrt{\mu_3} - \mu_3), \\ (-2\sqrt{\mu_4} + \mu_4 + 1, \sqrt{\mu_4} - \mu_4, \sqrt{\mu_4} - \mu_4, \mu_4), \\ (\mu_1, \mu_2, \mu_1(\mu_3 + \mu_4)/(\mu_1 + \mu_2), \mu_2(\mu_3 + \mu_4)/(\mu_1 + \mu_2)), \\ (\mu_1, \mu_1(\mu_2 + \mu_4)/(\mu_1 + \mu_3), \mu_3, \mu_3(\mu_2 + \mu_4)/(\mu_1 + \mu_3)), \\ (\mu_2(\mu_1 + \mu_3)/(\mu_2 + \mu_4), \mu_2, \mu_4(\mu_1 + \mu_3)/(\mu_2 + \mu_4), \mu_4), \\ (\mu_3(\mu_1 + \mu_2)/(\mu_3 + \mu_4), \mu_4(\mu_1 + \mu_2)/(\mu_3 + \mu_4), \mu_3, \mu_4). \end{cases}$$

The boundaries separating the various cases are given by the surfaces  $\{\mu \in \Delta_3 : \mu_1 - \mu_4 = 0, \mu_1 + \mu_2 \geq \sqrt{\mu_1}, \mu_1 + \mu_3 \geq \sqrt{\mu_1}\}$  and  $\{\mu \in \Delta_3 : \mu_2 - \mu_3 = 0, \mu_1 + \mu_2 \geq \sqrt{\mu_2}, \mu_2 + \mu_4 \geq \sqrt{\mu_2}\}$ .

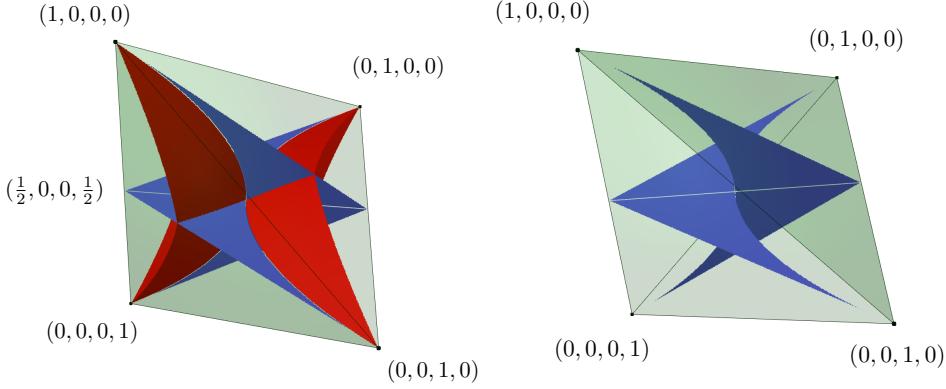


Fig. 5.2: The optimal value function of Theorem 5.10 subdivides the tetrahedron of probability distributions  $\mu$  (left). The surfaces that separate the various cases are shown in blue (right).

Theorem 5.10 involves a distinction into eight cases. This division of  $\Delta_3$  is shown in Figure 5.2. Each of the last four cases breaks into two subcases, since the numerator in the formulas is the absolute value of  $\mu_1\mu_4 - \mu_2\mu_3$ . The sign of this  $2 \times 2$  determinant matters for the pieces of our piecewise algebraic function. Thus, the tetrahedron  $\Delta_3$  is divided into 12 regions on which  $\mu \mapsto W_d(\mu, \mathcal{M})$  is algebraic. We now explain how to visualize Figure 5.2. The red surface consists of eight pieces that, together with the blue surface, separate the eight cases (this surface is not the model). Four convex regions are enclosed between the red surfaces and the edges they meet. These regions represent the first four cases in Theorem 5.10. For instance, the region containing the points  $(1, 0, 0, 0), (1/2, 0, 0, 1/2)$  corresponds to the first case. The remaining four regions are each bounded by two red and two blue pieces, and correspond to the last four cases. Each of these four regions is further split in two by the model. We do not depict this in our visualization. The two sides are determined by the sign of the determinant  $\mu_1\mu_4 - \mu_2\mu_3$ . The two blue surfaces in the right figure separate the various cases. These specify the points  $\mu \in \Delta_3$  with more than one optimal solution. For the proof of Theorem 5.10 and a simpler example see [46]. For further details we refer to [45].

### 5.3 Wasserstein meets Segre-Veronese

Returning to the general case, let  $\mathcal{M}$  be a smooth variety in  $\Delta_{n-1} \subset \mathbb{R}^n$ . For any  $\nu \in \Delta_{n-1}$ , we seek its distance to  $\mathcal{M}$  under our polyhedral norm. As before, the optimal point  $\nu^*$  determines a unique face  $F$  of the unit ball  $B = P_d^*$ . Given that face  $F$ , we now characterize optimality as in Lemma 5.2. Let  $\mathcal{F}$  be the set of all index pairs  $(i, j)$  such that the point  $\frac{1}{d_{ij}}(\mathbf{e}_i - \mathbf{e}_j)$  is a vertex and it lies in  $F$ . Let  $\ell_F$  be any linear functional on  $\mathbb{R}^n$  that attains its maximum over  $B$  at  $F$ . We work in the linear space spanned by the face:

$$L_F = \left\{ \sum_{(i,j) \in \mathcal{F}} \lambda_{ij} (\mathbf{e}_i - \mathbf{e}_j) : \lambda_{ij} \in \mathbb{R} \right\}. \quad (5.10)$$

The point  $\nu^*$  on  $\mathcal{M}$  that is closest to  $\mu$  is the solution of the following optimization problem:

$$\text{Minimize } \ell_F = \ell_F(\nu) \text{ subject to } \nu \in (\mu + L_F) \cap \mathcal{M}. \quad (5.11)$$

This is an optimization problem in the linear subspace  $L_F$ . With the notation in (5.10), the decision variables are  $\lambda_{ij}$  for  $(i, j) \in \mathcal{F}$ . The algebraic complexity of this problem is given by the polar degree (Theorem 5.5). The combinatorial complexity is governed by the facial structure of the Wasserstein ball  $B = P_d^*$  associated to a finite metric space  $([n], d)$ . We now focus on the polar dual, the  $(n-1)$ -dimensional Lipschitz polytope  $B^* = P_d$ . This polytope lives in  $\mathbb{R}^n / \mathbb{R}\mathbf{1} \simeq \mathbb{R}^{n-1}$ , and is defined in (5.6).

In the study of independence models  $\mathcal{M} \subset \Delta_{n-1}$ , the following metrics  $([n], d)$  arise:

- The discrete metric on any finite set  $[n]$  where  $d_{ij} = 1$  for distinct  $i, j$ .
- The  $L_0$ -metric on the Cartesian product  $[m_1] \times \cdots \times [m_k]$  where  $d_{ij} = \#\{l : i_l \neq j_l\}$ . Here  $i = (i_1, \dots, i_k)$  and  $j = (j_1, \dots, j_k)$  are elements in that Cartesian product.
- The  $L_1$ -metric on the Cartesian product  $[m_1] \times \cdots \times [m_k]$  where  $d_{ij} = \sum_{l=1}^k |i_l - j_l|$ .

For the last two metrics, the number of states of the relevant independence models is  $n = m_1 \cdots m_k$ . To compute Wasserstein distances, we need to describe the Lipschitz polytope  $P_d$  as explicitly as possible. All three metrics above are *graph metrics*. This means that there exists an undirected simple graph  $G$  with vertex set  $[n]$  such that  $d_{ij}$  is the length of the shortest path from  $i$  to  $j$  in  $G$ . The corresponding Wasserstein balls are called *symmetric edge polytopes*. They are investigated in [46, Section 4].

The following four independence models are used for the case studies in [46, Section 6]. We use the tuple  $((m_1)_{d_1}, \dots, (m_k)_{d_k})$  to denote the independence model with  $n = \prod_{i=1}^k \binom{m_i+d_i-1}{d_i}$  states where the  $i$ th entry  $(m_i)_{d_i}$  refers to a multinomial distribution with  $m_i$  possible outcomes and  $d_i$  trials, which can be interpreted as an unordered set of  $d_i$  identically distributed random variables on  $[m_i] = \{1, 2, \dots, m_i\}$ . The subscript  $d_i$  is omitted if  $d_i = 1$ . For example,  $(2_2, 2)$  is the independence model for three binary random variables where the first two are identically distributed. We list the  $n = 6$  states in the order 00, 10, 20, 01, 11, 21. These are the vertices of the associated graph  $G$ , which is the product of a 3-chain and a 2-chain. This model  $\mathcal{M}$  is the image of the map from the square  $[0, 1]^2$  into the simplex  $\Delta_5$  given by

$$(p, q) \mapsto (p^2 q, 2p(1-p)q, (1-p)^2 q, p^2(1-q), 2p(1-p)(1-q), (1-p)^2(1-q)). \quad (5.12)$$

**Example 5.11** We consider four models: the 3-bit model  $(2, 2, 2)$  with the  $L_0$ -metric on  $[2]^3$ , the model  $(3, 3)$  for two ternary variables with the  $L_1$ -metric on  $[3]^2$ , the model  $(2_6)$  for six identically distributed binary variables with the discrete metric on  $[7]$ , and the model  $(2_2, 2)$  in (5.12) with the  $L_1$ -metric on  $[3] \times [2]$ . In Table 5.1, we report the f-vectors of the Wasserstein balls for each of these models.

| $\mathcal{M}$ | $n$ | $\dim(\mathcal{M})$ | Metric $d$  | f-vector of the $(n-1)$ -polytope $P_d^*$ |
|---------------|-----|---------------------|-------------|---|
| $(2, 2, 2)$   | 8   | 3                   | $L_0 = L_1$ | (24, 192, 652, 1062, 848, 306, 38)        |
| $(3, 3)$      | 9   | 4                   | $L_1$       | (24, 216, 960, 2298, 3048, 2172, 736, 82) |
| $(2_6)$       | 7   | 1                   | discrete    | (42, 210, 490, 630, 434, 126)             |
| $(2_2, 2)$    | 6   | 2                   | $L_1$       | (14, 60, 102, 72, 18)                     |

Table 5.1: f-vectors of the Wasserstein balls for the four models in Example 5.11.

Independence models correspond in algebraic geometry to *Segre-Veronese varieties*. They are of considerable current interest in the study of tensor decompositions. We here replace the model, which is a semialgebraic set inside a simplex, by its complex Zariski closure in a projective space. This allows us to compute the algebraic degrees of our optimization problem.

The Segre-Veronese variety  $\mathcal{M} = ((m_1)_{d_1}, \dots, (m_k)_{d_k})$  is the embedding of  $\mathbb{P}^{m_1-1} \times \cdots \times \mathbb{P}^{m_k-1}$  in the projective space of partially symmetric tensors  $\mathbb{P}(\text{Sym}_{d_1} \mathbb{R}^{m_1} \otimes \cdots \otimes \text{Sym}_{d_k} \mathbb{R}^{m_k})$ . That projective space

equals  $\mathbb{P}^{n-1}$  where  $n = \prod_{i=1}^k \binom{m_i+d_i-1}{d_i}$ . By definition, the Segre-Veronese variety  $\mathcal{M}$  is the set of all tensors of rank one inside this projective space (tensors of rank one are discussed in more detail in Chapter 12).

**Example 5.12** Let  $k = 2$ . The Segre-Veronese variety  $\mathcal{M}((2)_2, (2)_1)$  is an embedding of  $\mathbb{P}^1 \times \mathbb{P}^1$  into  $\mathbb{P}^5$ , where it is a quartic surface. Its points are rank one tensors of format  $2 \times 2 \times 2$  which are symmetric in the first two indices. This model appears in the last row of Table 5.3.

We identify the projective variety  $\mathcal{M}$  with the intersection  $\mathcal{M} \cap \Delta_{n-1}$ . This is the set of real nonnegative points in  $\mathcal{M}$ . Thus, the independence model  $\mathcal{M}$  consists of nonnegative rank one tensors whose entries sum to 1. The dimension of  $\mathcal{M}$  is denoted  $\mathbf{m} := (m_1 - 1) + \dots + (m_k - 1)$ . The computation of the polar degrees of  $\mathcal{M}$  appears in the doctoral dissertation of Luca Sodomaco [165, Chapter 5]. We here state the result of this computation.

**Theorem 5.13 (Sodomaco)** *The polar degrees of the Segre-Veronese variety are*

$$\delta_{r-1}(\mathcal{M}) = \sum_{s=0}^{\mathbf{m}-n+1+r} (-1)^s \binom{\mathbf{m}-s+1}{n-r} (\mathbf{m}-s)! \left( \sum_{i_1+\dots+i_k=s} \prod_{l=1}^k \frac{\binom{m_l}{i_l} d_l^{m_l-1-i_l}}{(m_l-1-i_l)!} \right). \quad (5.13)$$

Here  $r$  is any integer in the range  $n - 1 - \dim(\mathcal{M}) \leq r \leq \dim(\mathcal{M}^*)$ .

We next examine this formula for various special cases starting with the binary case.

**Corollary 5.14** *Let  $\mathcal{M}$  be the  $k$ -bit independence model. The formula (5.13) specializes to*

$$\delta_{r-1}(\mathcal{M}) = \sum_{s=0}^{k-2^k+1+r} (-1)^s \binom{k+1-s}{2^k-r} (k-s)! 2^s \binom{k}{s}. \quad (5.14)$$

In algebraic geometry language, our model  $\mathcal{M}$  here is the Segre embedding of  $(\mathbb{P}^1)^k$  into  $\mathbb{P}^{2^k-1}$ . This is the toric variety associated with the  $k$ -cube, so its degree is the normalized volume of the cube, which is  $k!$ . The polar degrees  $\delta_{r-1}$  in (5.14) are shown for  $k \leq 7$  in Table 5.2. The indices  $r$  with  $\delta_{r-1} \neq 0$  range from  $\text{codim}(\mathcal{M}) = 2^k - 1 - k$  to  $\dim(\mathcal{M}^*) = 2^k - 1$ . For the sake of the table's layout, we shift the indices on each row so that the row labeled 0 contains  $\delta_{\text{codim}(\mathcal{M})-1} = \text{degree}(\mathcal{M}) = k!$ . The dual variety  $\mathcal{M}^*$  is a hypersurface of degree  $\delta_{2^k-2}$  known as the *hyperdeterminant* of format  $2^k$ . For instance, for  $k = 3$ , this hypersurface in  $\mathbb{P}^7$  is the  $2 \times 2 \times 2$ -hyperdeterminant which has degree four. The entries in the first column ( $k = 2$ ) corresponds to the three scenarios in Figure 5.1, where the algebraic degree equals 2.

| $r - \text{codim}(\mathcal{M})$ | $k = 2$ | $k = 3$ | $k = 4$ | $k = 5$ | $k = 6$ | $k = 7$ |
|---------------------------------|---------|---------|---------|---------|---------|---------|
| 0                               | 2       | 6       | 24      | 120     | 720     | 5040    |
| 1                               | 2       | 12      | 72      | 480     | 3600    | 30240   |
| 2                               | 2       | 12      | 96      | 840     | 7920    | 80640   |
| 3                               |         | 4       | 64      | 800     | 9840    | 124320  |
| 4                               |         |         | 24      | 440     | 7440    | 120960  |
| 5                               |         |         |         | 128     | 3408    | 75936   |
| 6                               |         |         |         |         | 880     | 30016   |
| 7                               |         |         |         |         |         | 6816    |

Table 5.2: The polar degrees  $\delta_{r-1}(\mathcal{M})$  of the  $k$ -bit independence model for  $k \leq 7$ .

We briefly discuss the independence models  $(m_1, m_2)$  for two random variables. These are the classical contingency tables of format  $m_1 \times m_2$ . Here,  $n = m_1 m_2$  and  $\mathbf{m} = m_1 + m_2 - 2$ . The Segre variety  $\mathcal{M} = \mathbb{P}^{m_1-1} \times \mathbb{P}^{m_2-1} \subset \mathbb{P}^{n-1}$  consists of  $m_1 \times m_2$  matrices of rank one.

**Corollary 5.15** *The Segre variety of  $m_1 \times m_2$  matrices of rank one has the polar degrees*

$$\delta_{r-1}(\mathcal{M}) = \sum_{s=0}^{\mathbf{m}-n+1+r} (-1)^s \binom{\mathbf{m}-s+1}{n-r} (\mathbf{m}-s)! \left( \sum_{i+j=s} \frac{\binom{m_1}{i}}{(m_1-1-i)!} \cdot \frac{\binom{m_2}{j}}{(m_2-1-j)!} \right).$$

The polar degrees above serve as upper bounds for any particular Wasserstein distance problem. For a fixed model  $\mathcal{M}$ , the equality in Theorem 5.5 holds only when the data  $(\ell, L)$  is generic. However, for the optimization problem in (5.11), the linear space  $L = L_F$  and the linear functional  $\ell = \ell_F$  are very specific. They depend on the Lipschitz polytope  $P_d$  and the type  $F$  of the optimal solution  $v^*$ . For such specific scenarios, we only get an inequality.

**Proposition 5.16** *Consider the problem (5.11) for the independence model  $((m_1)_{d_1}, \dots, (m_k)_{d_k})$  with a given face  $F$  of the Wasserstein ball  $B = P_d^*$ . The degree of the optimal solution  $v^*$  as an algebraic function of the data  $\mu$  is bounded above by the polar degree  $\delta_{r-1}$  in (5.13).*

**Proof** This follows from Theorem 5.5. The upper bound relies on general principles of algebraic geometry. Namely, the graph of the map  $\mu \mapsto v^*(\mu)$  is an irreducible variety, and we seek its degree over  $\mu$ . The map depends on the parameters  $(\ell, L)$ . When the coordinates of  $L$  and  $\ell$  are independent transcendentals then the algebraic degree is the polar degree  $\delta_{r-1}$ . That algebraic degree can only go down when these coordinates take on special values in the real numbers. That same semi-continuity argument holds for most polynomial optimization problems, including the Euclidean distance optimization in the last lecture.  $\square$

We now examine the drop in algebraic degree for the four models in Example 5.11. In the language of algebraic geometry, they are the Segre threefold  $\mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1$  in  $\mathbb{P}^7$ , the variety  $\mathbb{P}^2 \times \mathbb{P}^2$  of rank one  $3 \times 3$  matrices in  $\mathbb{P}^8$ , the rational normal curve  $\mathbb{P}^1$  in  $\mathbb{P}^6 = \mathbb{P}(\text{Sym}_6(\mathbb{R}^2))$ , and the Segre-Veronese surface  $\mathbb{P}^1 \times \mathbb{P}^1$  in  $\mathbb{P}^5 = \mathbb{P}(\text{Sym}_2(\mathbb{R}^2) \times \text{Sym}_1(\mathbb{R}^2))$ . The finite metrics  $d$  are specified in the fourth column of Table 5.1. The fifth column in Table 5.1 records the combinatorial complexity of our optimization problem, while the algebraic complexity is recorded in Table 5.3.

| $\mathcal{M}$ | Polar degrees                | Maximal degree             | Average degree                           |
|---------------|------------------------------|----------------------------|--|
| $(2, 2, 2)$   | $(0, 0, 0, 6, 12, 12, 4)$    | $(0, 0, 0, 4, 12, 6, 0)$   | $(0, 0, 0, 2.138, 6.382, 3.8, 0)$        |
| $(3, 3)$      | $(0, 0, 0, 6, 12, 12, 6, 3)$ | $(0, 0, 0, 2, 8, 6, 6, 0)$ | $(0, 0, 0, 1.093, 3.100, 4.471, 6.0, 0)$ |
| $(2_6)$       | $(0, 0, 0, 0, 6, 10)$        | $(0, 0, 0, 0, 6, 5)$       | $(0, 0, 0, 0, 6, 5)$                     |
| $(2_2, 2)$    | $(0, 0, 4, 6, 4)$            | $(0, 0, 3, 5, 2)$          | $(0, 0, 2.293, 3.822, 2.0)$              |

Table 5.3: The algebraic degrees of the problem (5.8) for the four models in Example 5.11.

The second column in Table 5.3 gives the vector  $(\delta_0, \delta_1, \dots, \delta_{n-2})$  of polar degrees. The third and fourth column are the results of a computational experiment. For each model, we take 1000 uniform samples  $\mu$  with rational coordinates from  $\Delta_{n-1}$ , and we solve the optimization problem (5.8). The output is an exact representation of the optimal solution  $v^*$ . This includes the optimal face  $F$  that specifies  $v^*$ , along with its maximal ideal over  $\mathbb{Q}$ . The algebraic degree of the optimal solution  $v^*$  is computed as the number of complex zeros of that maximal ideal. This number is bounded above by the polar degree (cf. Proposition 5.16). The fourth column in Table 5.3 shows the average of the algebraic degrees we found. For example, for the 3-bit model  $(2, 2, 2)$  we have  $\delta_3 = 6$ , corresponding to  $P_d^*$  touching  $\mathcal{M}$  at a

3-face  $F$ . However, the maximum degree we saw in our computations was 4, with an average degree of 2.138. For 4-faces  $F$ , we have  $\delta_4 = 12$ , and this degree was attained in some runs. The average was 6.382.

Such computational experiments are organized naturally into three stages: (1) combinatorial preprocessing, (2) numerical optimization, and (3) algebraic postprocessing. Our object of interest is a model  $\mathcal{M}$  in the simplex  $\Delta_{n-1}$ , typically one of the independence models  $((m_1)_{d_1}, \dots, (m_k)_{d_k})$  where  $n = \prod_{i=1}^k \binom{m_i+d_i-1}{d_i}$ .

The state space  $[n]$  is a metric space, with metric given by the matrix  $d = (d_{ij})$ . This matrix of pairwise distances is part of our input. It defines the Lipschitz polytope  $P_d$  and its dual, the Wasserstein ball  $P_d^*$ . Our first algorithm computes these combinatorial objects.

---

**Algorithm 1:** Combinatorial Preprocessing

---

- 1 **Input:** An  $n \times n$  symmetric matrix  $d = (d_{ij})$ .
  - 2 **Output:** A description of all facets  $F$  of the Wasserstein ball  $P_d^*$ .
  - 3 From the inequality presentation in (5.6), find all vertices of the Lipschitz polytope  $P_d$ . These vertices are the inner normal vectors  $\ell_F$  to the facets  $F$  of  $P_d^*$ . Store them.
  - 4 Determine an inequality description of the cone  $C_F$  over each facet  $F$ .
  - 5 **return** the list of pairs  $(\ell_F, C_F)$ , one for each vertex of the Lipschitz polytope  $P_d$ .
- 

In the original study [46], the software **Polymake** was used to run Algorithm 1. We next solve the optimization problem in (5.8), by examining each facet  $F$  of the Wasserstein ball. The problem is that in (5.11) but with the linear space  $L_F$  now replaced by the convex cone  $C_F$  that is spanned by  $F$ .

---

**Algorithm 2:** Numerical Optimization

---

- 1 **Input:** Model  $\mathcal{M}$  and a point  $\mu$  in the simplex  $\Delta_{n-1}$ ; complete output from Algorithm 1.
  - 2 **Output:** The optimal solution  $v^*$  in (5.8) along with its type  $G$ .
  - 3 **for** each facet  $F$  of the Wasserstein ball  $P_d^*$  **do**
  - 4     Apply global optimization methods to identify a point  $v^* \in \mathcal{M}$  that minimizes  $\ell_F = \ell_F(v)$  subject to  $v \in (\mu + C_F) \cap \mathcal{M}$ .
  - 5     Identify the unique face  $G$  of  $F$  whose span contains  $v^*$  in its relative interior.
  - 6     Identify a basis of vectors  $e_i - e_j \in C_G$  for the linear space  $L_G$  spanned by  $G$ .
  - 7     Store the optimal solution  $v^*$  and a basis for the linear subspace  $L_G$  of  $\mathbb{R}^n$ .
  - 8 **end**
  - 9 Among all candidate solutions, identify the solution  $v^*$  for which the Wasserstein distance  $W_d(\mu, v^*)$  to the given data point  $\mu$  is smallest. Record its type  $G$ .
  - 10 **return** The optimal solution  $v^*$ , its associated linear space  $L_G$ , and the facet normal  $\ell_G$ .
- 

In [46], the software **SCIP** was used to run Algorithm 2. **SCIP** employs sophisticated branch-and-cut strategies to solve constrained polynomial optimization problems via LP relaxation. Algorithm 1 is guaranteed to find the global optimum for our problem (5.8). Moreover, it furnishes an identification of the combinatorial type. This serves as the input to the symbolic computation in Algorithm 3 below.

Algorithm 3 can be carried out with a computer algebra system like **Macaulay2**. Steps 2 and 4 are the result of standard Gröbner basis calculations. The pipeline is illustrated with examples in [46, Section 6].

**Algorithm 3:** Algebraic Postprocessing

---

1 **Input:** The optimal solution  $(\nu^*, G)$  to (5.8) in the form found by Algorithm 2.  
 2 **Output:** The maximal ideal in the polynomial ring  $\mathbb{Q}[\nu_1, \dots, \nu_n]$  which has the zero  $\nu^*$ .  
 3 Use Lagrange multipliers to give polynomial equations that characterize the critical points of the linear function  $\ell_F$  on the subvariety  $(\mu + L_G) \cap \mathcal{M}$  in the affine space  $\mathbb{R}^n$ .  
 4 Eliminate all variables representing Lagrange multipliers from the ideal in the previous step. This ideal lives in  $\mathbb{Q}[\nu_1, \dots, \nu_n]$ .  
 5 **if** the ideal in step 4 is maximal **then**  
   6   | Call the ideal  $M$ .  
   7 **else**  
     8   | Remove extraneous primary components to get the maximal ideal  $M$  of  $\nu^*$ .  
   9 **end**  
 10 Determine the degree of  $\nu^*$ , which is the dimension of  $\mathbb{Q}[\nu_1, \dots, \nu_n]/M$  over  $\mathbb{Q}$ .  
 11 **return** the generators for the maximal ideal  $M$  along with the degree found in Step 10.

---

## **Chapter 6**

## **Curvature**

## 6.1 Plane Curves

In this section, we consider a smooth algebraic curve  $C \subset \mathbb{R}^2$  given as the zero set of an irreducible polynomial of degree  $d \geq 1$ :

$$f(x_1, x_2) \in \mathbb{R}[x_1, x_2]$$

Let  $\mathbf{x} = (x_1, x_2)$ . Geometrically, the *curvature* of  $C$  at a point  $\mathbf{x} \in C$  is the rate of change at  $\mathbf{x}$  of a unit normal vector traveling along  $C$ . To be precise, define

$$N(\mathbf{x}) := \frac{1}{\|\nabla f(\mathbf{x})\|} \nabla f(\mathbf{x}), \quad (6.1)$$

where  $\nabla f(\mathbf{x}) = (\partial f / \partial x_1, \partial f / \partial x_2)$  is the gradient of  $f$ . Then, for all  $\mathbf{x} \in C$ ,  $N(\mathbf{x})$  returns a normal vector of  $C$  at  $\mathbf{x}$ ; one calls  $N(\mathbf{x})$  a *unit normal field*. Similarly, a *unit tangent field* is given by

$$T(\mathbf{x}) := (N(\mathbf{x})_2, -N(\mathbf{x})_1).$$

The curvature of  $C$  at  $\mathbf{x}$  is defined as the (signed) magnitude of the derivative of  $N(\mathbf{x})$  in tangent direction:

$$c(\mathbf{x}) := \left\langle T(\mathbf{x}), T(\mathbf{x})_1 \cdot \frac{\partial N}{\partial x_1}(\mathbf{x}) + T(\mathbf{x})_2 \cdot \frac{\partial N}{\partial x_2}(\mathbf{x}) \right\rangle. \quad (6.2)$$

Since the derivative of a unit normal field at a curve always points in tangent direction, (6.2) has the equivalent formulation  $T(\mathbf{x})_1 \cdot \frac{\partial N}{\partial x_1}(\mathbf{x}) + T(\mathbf{x})_2 \cdot \frac{\partial N}{\partial x_2}(\mathbf{x}) = c(\mathbf{x}) \cdot T(\mathbf{x})$ .

**Example 6.1** Consider the ellipse  $f(\mathbf{x}) = x_1^2 + \frac{1}{4}x_2^2 - 1$ . The left picture in Figure 6.1 shows the ellipse in green and the normal field  $N(\mathbf{x})$  in yellow. The right picture displays the curvature via  $c(\mathbf{x}) \cdot T(\mathbf{x})$ . The magnitude of a yellow vector attached to a point  $\mathbf{x}$  in the right picture gives the curvature at  $\mathbf{x}$ . On the top and bottom, where the ellipse is rather flat, the normal vectors do not change much, hence the curvature is small. On the sides, normal vectors change more rapidly, so there the curvature is larger  $\diamond$

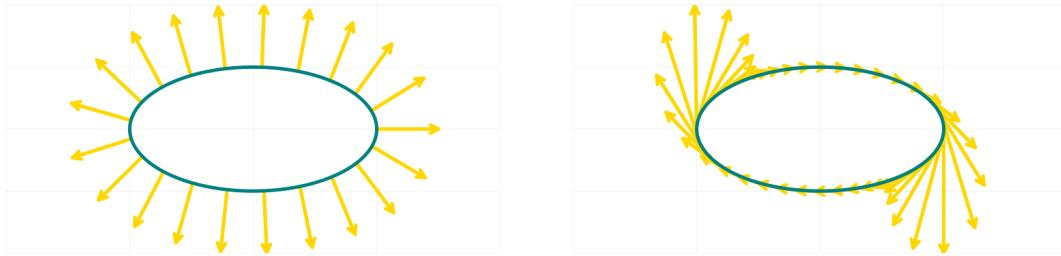


Fig. 6.1: The left picture shows a unit normal field of the ellipse  $x_1^2 + \frac{1}{4}x_2^2 - 1 = 0$ , and the right picture shows how the normal field changes when travelling along the ellipse.

The inverse of the signed curvature  $r(\mathbf{x}) := c(\mathbf{x})^{-1}$  is called the (signed) *radius of curvature*. This is because  $C$  contains an infinitesimally small arc of a circle with radius  $|r(\mathbf{x})|$  and center  $\mathbf{x} - r(\mathbf{x}) \cdot N(\mathbf{x})$ . This center is also called a *focal point* or *center of curvature* of  $C$ . The reason for the negative sign in this formula is that a normal vector pointing towards the focal point changes towards the direction that is opposite to  $T(\mathbf{x})$ ; see Figure 6.2 for an illustration.

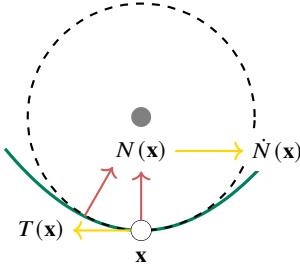


Fig. 6.2: The picture shows a green curve that contains an infinitesimally small arc of the dashed circle. The grey center of the circle is a focal point of the green curve. The red vertical normal vector  $N(\mathbf{x})$  pointing towards the focal point changes into a normal vector that is slightly tilted in the direction opposite to  $T(\mathbf{x})$ . This means that  $\dot{N}(\mathbf{x}) := T(\mathbf{x})_1 \frac{\partial N(\mathbf{x})}{\partial x_1} + T(\mathbf{x})_2 \frac{\partial N(\mathbf{x})}{\partial x_2}$  is a negative multiple of  $T(\mathbf{x})$ .

**Proposition 6.2** *The Zariski closure of the set of all centers of curvature of a plane curve  $C$  is the evolute of  $C$ , as defined in Section 1.3.*

**Proof** We consider a local parametrization  $\gamma(t)$  of  $C$  with  $\gamma(0) = \mathbf{x}$  and  $\dot{\gamma}(0) = T(\mathbf{x})$ . Then,  $\varepsilon(t) := \gamma(t) - r(\gamma(t)) \cdot N(\gamma(t))$  gives a local parametrization of the curve  $E$  that is traced by the centers of curvature of  $C$ . We have

$$\dot{\varepsilon}(0) = T(\mathbf{x}) - \langle \nabla r(\mathbf{x}), T(\mathbf{x}) \rangle \cdot N(\mathbf{x}) - r(\mathbf{x}) \cdot (T(\mathbf{x})_1 \cdot \frac{\partial N}{\partial x_1}(\mathbf{x}) + T(\mathbf{x})_2 \cdot \frac{\partial N}{\partial x_2}(\mathbf{x})).$$

Since  $T(\mathbf{x})_1 \cdot \frac{\partial N}{\partial x_1}(\mathbf{x}) + T(\mathbf{x})_2 \cdot \frac{\partial N}{\partial x_2}(\mathbf{x}) = c(\mathbf{x}) \cdot T(\mathbf{x}) = r(\mathbf{x})^{-1} \cdot T(\mathbf{x})$ , we get

$$\dot{\varepsilon}(0) = -\langle \nabla r(\mathbf{x}), T(\mathbf{x}) \rangle \cdot N(\mathbf{x}). \quad (6.3)$$

This means that the tangent line of  $E$  at  $\varepsilon(0) = \mathbf{x} - r(\mathbf{x}) \cdot N(\mathbf{x})$  is the normal line of  $C$  at  $\mathbf{x}$ . Hence, the curve  $E$  is the envelope of the normal lines of  $C$ ; in other words, the evolute of  $C$ .  $\square$

The previous proof shows in particular that the absolute value of the curvature  $|c(\mathbf{x})|$  at a point  $\mathbf{x} \in C$  is the inverse distance from  $\mathbf{x}$  to its corresponding point on the evolute. Indeed, the latter point is  $\mathbf{x} - r(\mathbf{x}) \cdot N(\mathbf{x})$ , and so its distance from  $\mathbf{x}$  is  $|r(\mathbf{x})| = |c(\mathbf{x})|^{-1}$ .

In the remainder of this section, we will study two types of points: *inflection points* and *points of critical curvature*. These points exhibit special curvature of  $C$ . Inflection points are points where  $C$  is locally flat, and critical curvature points are points where the curvature has a local extremum.

**Definition 6.3** Let  $\mathbf{x} \in C$ .

1. We call  $\mathbf{x}$  an *inflection point* if  $c(\mathbf{x}) = 0$ .
2. We call  $\mathbf{x}$  a *critical curvature point* if  $\mathbf{x}$  is a critical point of the function  $C \rightarrow \mathbb{R}$ ,  $x \mapsto c(x)$ .

**Example 6.4** We consider the Trott curve  $f(\mathbf{x}) = 144(x^4 + y^4) - 225(x^2 + y^2) + 350x^2y^2 + 81$  as we did in Figure 2.1. The Trott curve has degree  $d = 4$ . Figure 6.3 shows the curve in green. We first compute inflection points using `HomotopyContinuation.jl` [29] and the formulation in Lemma 6.7.

```
using HomotopyContinuation, LinearAlgebra
@var x y z
v = [x; y; z]
F = 144*(x^4 + y^4) - 225*(x^2 + y^2) + 350*x^2*y^2 + 81*z^4
dF = differentiate(F, v)
H0 = differentiate(dF, v)
```

```
f = subs(F, z=>1)
h = subs(-det(H0), z=>1)
inflection_points = solve([f; h])
```

By Theorem 6.8, there are  $3d(d - 2) = 24$  complex inflection points. Klein [112] proved that at most  $d(d - 2) = 8$  can be real and indeed, in this case, we find 8 real inflection points. They are the yellow points in Figure 6.3. Next, we compute critical curvature points using the equations in Lemma 6.10.

```
f1, f2 = dF[1:2]
f11, f12, f12, f22 = H0[1:2,1:2]
hx, hy = differentiate(h,[x; y])
g = f1 * f2 * (f11-f22) + f12 * (f2^2- f1^2)
c = subs(f2 * hy - f1 * hx - 3 * h * g, z=>1)
crit_curv = solve([f; c])
```

By Theorem 6.11, there are  $2d(3d - 5) = 56$  complex critical curvature points. We find that 24 of them are real; out of these 8 are close (but not equal to) to the 8 inflection points, which is why they are not visible in the picture. They are shown as red points in Figure 6.3. By Proposition 6.5, the critical curvature points correspond to the cusps on the evolute, which is illustrated in Figure 6.4.  $\diamond$

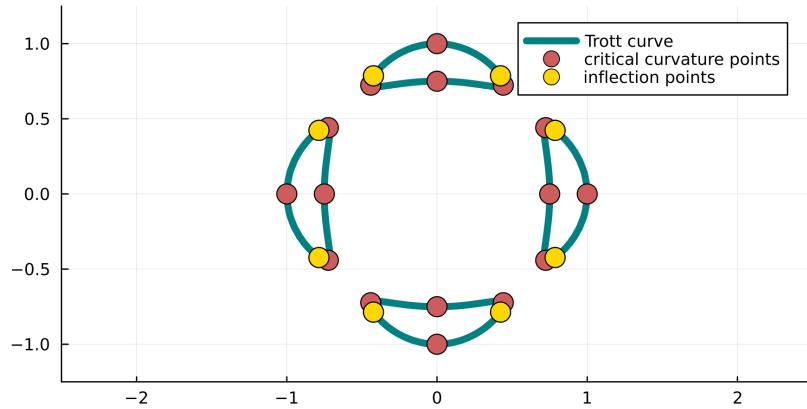


Fig. 6.3: The Trott curve  $144(x^4 + y^4) - 225(x^2 + y^2) + 350x^2y^2 + 81 = 0$  in green together with its inflection points (yellow points) and critical curvature points (red points).

**Proposition 6.5** Let  $C \subset \mathbb{R}^2$  be a smooth algebraic curve and  $E \subset \mathbb{R}^2$  be its evolute. For  $\mathbf{x} \in C$ , let  $r(\mathbf{x})$  be the radius of curvature and  $\Gamma(\mathbf{x}) := \mathbf{x} - r(\mathbf{x}) \cdot N(\mathbf{x}) \in E$  be the corresponding point on the evolute.

1.  $\mathbf{x}$  is an inflection point if and only if  $\Gamma(\mathbf{x})$  is a point at infinity.
2.  $\mathbf{x}$  is a point of critical curvature if and only if  $E$  has a cusp at  $\Gamma(\mathbf{x})$ .

**Proof** Recall that  $c(\mathbf{x}) = r(\mathbf{x})^{-1}$ . The point  $\Gamma(\mathbf{x})$  is at infinity if and only if  $c(\mathbf{x}) = 0$ , which means that  $\mathbf{x}$  is an inflection point. This proves the first item. For the second item, we consider a local parametrization  $\gamma(t)$  of  $C$  with  $\gamma(0) = \mathbf{x}$  and  $\dot{\gamma}(0) = T(\mathbf{x})$ . As in the proof of Proposition 6.2,  $\varepsilon(t) := \Gamma(\gamma(t))$  gives a

local parametrization of the evolute  $E$ . The evolute has a cusp at  $\Gamma(\mathbf{x}) = \varepsilon(0)$  if and only if  $\dot{\varepsilon}(0) = 0$ . By (6.3), the latter is equivalent to  $\langle \nabla r(\mathbf{x}), T(\mathbf{x}) \rangle = 0$ . The latter equation holds exactly when the curvature  $c(\mathbf{x})$  is critical at  $\mathbf{x}$  since  $\nabla c(\mathbf{x}) = \frac{-\nabla r(\mathbf{x})}{r(\mathbf{x})^2}$ .  $\square$

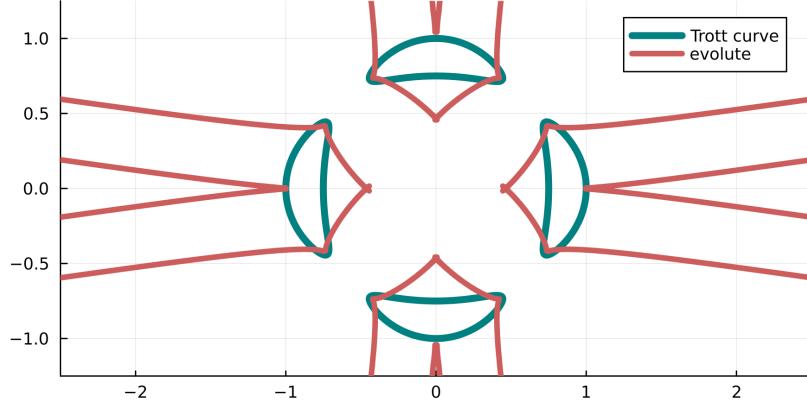


Fig. 6.4: The picture shows the Trott curve  $144(x^4 + y^4) - 225(x^2 + y^2) + 350x^2y^2 + 81 = 0$  in green and its evolute in red. The cusps on the evolute correspond to the red critical curvature points in Figure 6.3. The Trott curve has 24 real critical curvature points. Out of those 8 have a radius of curvature which exceeds the boundary of this picture. This is why we only see 16 cusps. We thank Emil Horobet and Pierpaola Santarsiero for helping us computing the equation for the evolute.

We want to count the number of complex inflection and critical curvature points for a curve given by a general polynomial  $f$ . For this, let us first understand the curvature  $c(\mathbf{x})$  better. In the following, we denote partial derivatives by  $f_i := \frac{\partial f}{\partial x_i}$ ,  $f_{i,j} := \frac{\partial^2 f}{\partial x_i \partial x_j}$  and the *Hessian* by

$$H := \begin{pmatrix} f_{1,1} & f_{1,2} \\ f_{1,2} & f_{2,2} \end{pmatrix}.$$

**Lemma 6.6** *The curvature equals*  $c(\mathbf{x}) = \frac{1}{\|\nabla f(\mathbf{x})\|} \cdot T(\mathbf{x})^\top H(\mathbf{x}) T(\mathbf{x})$

**Proof** Recall that  $c(\mathbf{x}) = \langle T(\mathbf{x}), T(\mathbf{x})_1 \cdot \frac{\partial N}{\partial x_1}(\mathbf{x}) + T(\mathbf{x})_2 \cdot \frac{\partial N}{\partial x_2}(\mathbf{x}) \rangle$ . By applying the product rule on (6.1), we obtain

$$T(\mathbf{x})_1 \cdot \frac{\partial N}{\partial x_1}(\mathbf{x}) + T(\mathbf{x})_2 \cdot \frac{\partial N}{\partial x_2}(\mathbf{x}) = \|\nabla f(\mathbf{x})\|^{-1} \cdot H(\mathbf{x}) T(\mathbf{x}) + a(\mathbf{x}) \cdot \nabla f(\mathbf{x})$$

for some scalar function  $a(\mathbf{x})$ . Since  $\langle T(\mathbf{x}), \nabla f(\mathbf{x}) \rangle = 0$ , this gives the asserted formula.  $\square$

We can write the formula from Lemma 6.6 more explicitly as

$$c(\mathbf{x}) = \frac{f_{1,1} \cdot f_{2,2}^2 - 2f_{1,2} \cdot f_1 \cdot f_2 + f_{2,2} \cdot f_1^2}{(f_1^2 + f_2^2)^{\frac{3}{2}}}(\mathbf{x}). \quad (6.4)$$

For what follows it will be helpful to embed  $C$  into complex projective space. We write

$$F(x_0, x_1, x_2) := x_0^d f\left(\frac{x_1}{x_0}, \frac{x_2}{x_0}\right)$$

for the homogenization of  $f$ . Moreover, let

$$H_0 = \begin{pmatrix} F_{0,0} & F_{0,1} & F_{0,2} \\ F_{0,1} & F_{1,1} & F_{1,2} \\ F_{0,2} & F_{1,2} & F_{2,2} \end{pmatrix}.$$

denote the Hessian of  $F$  viewed as polynomial in  $\mathbf{x}$  by setting  $x_0 = 1$ . We can rewrite the curvature of  $C$  in terms of  $F$ . The next lemma goes back to Salmon [161].

**Lemma 6.7** *The curvature at  $\mathbf{x} \in C$  can be expressed as*

$$c(\mathbf{x}) = \frac{-\det H_0}{(d-1)^2 \cdot (f_1^2 + f_2^2)^{\frac{3}{2}}}(\mathbf{x}).$$

**Proof** Homogenizing polynomials in (6.4), we get  $c = \frac{P}{Q}$ , where  $P = F_{1,1} \cdot F_2^2 - 2F_{1,2} \cdot F_1 \cdot F_2 + F_{2,2} \cdot F_1^2$  and  $Q = (F_1^2 + F_2^2)^{\frac{3}{2}}$ . By Euler's formula for homogeneous functions, we have

$$(d-1) \cdot F_j = x_0 F_{0,j} + x_1 F_{1,j} + x_2 F_{2,j}, \quad 0 \leq j \leq 2. \quad (6.5)$$

Substituting the  $F_j$  in  $P$  gives

$$\begin{aligned} (d-1)^2 \cdot P &= (F_{1,1} F_{2,2} - F_{1,2}^2) \cdot (x_1^2 F_{1,1} + x_2^2 F_{2,2} + 2(x_0 x_1 F_{0,1} + x_0 x_2 F_{0,2} + x_1 x_2 F_{1,2})) \\ &\quad + x_0^2 (F_{0,1}^2 F_{2,2} - 2F_{1,2} F_{0,1} F_{0,2} + F_{1,1} F_{0,2}^2). \end{aligned}$$

Furthermore, we have  $0 = dF = x_0 F_0 + x_1 F_1 + x_2 F_2$  on  $C$ . Substituting (6.5) into this equation, we have

$$0 = x_0^2 F_{0,0} + x_1^2 F_{1,1} + x_2^2 F_{2,2} + 2(x_0 x_1 F_{0,1} + x_0 x_2 F_{0,2} + x_1 x_2 F_{1,2})$$

on  $C$ . We obtain

$$P = \frac{x_0^2}{(d-1)^2} \cdot (-(F_{1,1} F_{2,2} - F_{1,2}^2) F_{0,0} + F_{0,1}^2 F_{2,2} - 2F_{1,2} F_{0,1} F_{0,2} + F_{1,1} F_{0,2}^2) = \frac{-x_0^2 \cdot \det H_0}{(d-1)^2}.$$

Setting  $x_0 = 1$  finishes the proof.  $\square$

We can now count the inflection points of a general curve.

**Theorem 6.8** *The number of complex inflection points of a curve  $C$  defined by a general polynomial  $f$  of degree  $d$  is  $3d(d-2)$ .*

This theorem was first proved by Klein [112]. He also proved that the number of real inflection points is at most one third; i.e.,  $d(d-2)$ . Here, we give a short proof for the complex count.

**Proof** By Lemma 6.7, inflection points on  $C$  are given as the zero set  $f = \det H_0 = 0$ , which is a system of two equations in two variables  $\mathbf{x} = (x_1, x_2)$ . The degree of  $f$  is  $d$  and the degree of  $\det H_0$  is  $3(d-2)$ . Bézout's theorem implies that the number of inflection points is at most  $3d(d-2)$ . To show that the number is also at least  $3d(d-2)$ , we show that there exist polynomials with this number of inflection points.

Consider a univariate polynomial  $g(x_1) \in \mathbb{R}[x_1]$  of degree  $d$  and let  $G(x_0, x_1)$  be its homogenization. Let also  $M := \begin{pmatrix} G_{0,0} & G_{0,1} \\ G_{0,1} & G_{1,1} \end{pmatrix}$  be the Hessian of  $G$  with respect to the variables  $x_0, x_1$ . We assume (1) that  $g$  has  $d$  regular zeros, (2) that  $\det M = 0$  has only regular zeros, and (3) that  $g = \det M = 0$  has no solutions. All three are Zariski open conditions, so almost all polynomials  $g$  satisfy this assumption. Define

$$f(x_1, x_2) := x_2^d - g(x_1).$$

Observe that in this case  $\det H_0 = d(d-1) \cdot x_2^{d-2} \cdot \det M$ . Consequently,  $\det H_0 = 0$  if and only if either  $x_2 = 0$ , or  $x_1$  is among the  $2(d-2)$  zeros of  $\det M$ . This means that  $\mathbf{x}$  is an inflection point if either  $\mathbf{x} = (x_1, 0)$  and  $x_1$  is a zero of  $g$ , or  $\mathbf{x} = (x_1, x_2)$  where  $x_1$  is a zero of  $\det M$  and  $x_2^d = g(x_1)$ . In the first case, we find  $d$  inflection points and each has multiplicity  $d-2$ . In the second case, since  $g(x_1) \neq 0$ , we find  $2d(d-2)$  many regular inflection points with multiplicity one. Now, if we perturb  $f$  slightly, the  $d$  points with multiplicity will split into  $d(d-2)$  inflection points, while the other  $2d(d-2)$  inflection points will remain distinct. In total, this gives  $3d(d-2)$  inflection points.  $\square$

**Corollary 6.9** *For a plane curve  $C$  defined by a general polynomial  $f$  of degree  $d$ , its evolute has degree  $3d(d-1)$ .*

**Proof** We compute the desired degree by intersecting the evolute with the line at infinity. For that, we consider the Zariski closure  $\bar{C}$  of the curve  $C$  in the complex projective plane  $\mathbb{P}_{\mathbb{C}}^2$ . By Proposition 6.2, the evolute is the image of  $\Gamma : \bar{C} \rightarrow \mathbb{P}_{\mathbb{C}}^2, \mathbf{x} \mapsto \mathbf{x} - r(\mathbf{x}) \cdot N(\mathbf{x})$ . A point  $\Gamma(\mathbf{x})$  on the evolute can be at infinity for two reasons: Either  $\mathbf{x} \in \bar{C}$  is at infinity, or  $\mathbf{x}$  is a finite point and a complex inflection point of the curve  $C$  by Proposition 6.5. Since  $C$  is a general curve of degree  $d$ , there are  $d$  points  $\mathbf{x}$  of the first kind and  $3d(d-2)$  points  $\mathbf{x}$  of the second kind by Theorem 6.8. For each of the  $d$  points  $\mathbf{x} \in \bar{C}$  at infinity, Salmon [161, §119] shows that  $\Gamma(\mathbf{x})$  is a cusp whose tangent line is the line at infinity. Hence, when intersecting the evolute with the line at infinity, there are  $d$  cusps (that count with multiplicity three each) plus  $3d(d-2)$  points that correspond to the complex inflection points of  $C$ . All in all, the degree of the evolute is  $3d + 3d(d-2) = 3d(d-1)$ .  $\square$

Let us now find polynomial equations for critical curvature.

**Lemma 6.10** *Critical curvature points on the curve  $C = \{f(\mathbf{x}) = 0\}$  are defined by the equation*

$$(f_1^2 + f_2^2) \cdot \left( f_2 \cdot \frac{\partial \det H_0}{\partial x_1} - f_1 \cdot \frac{\partial \det H_0}{\partial x_2} \right) - 3 \det H_0 \cdot g = 0,$$

where  $g = f_1 f_2 \cdot (f_{1,1} - f_{2,2}) + f_{1,2}(f_2^2 - f_1^2)$ .

**Proof** Critical curvature points on  $C$  are defined by the equations  $f = 0$  and  $f_2 \cdot \frac{\partial c(\mathbf{x})}{\partial x_1} - f_1 \cdot \frac{\partial c(\mathbf{x})}{\partial x_2} = 0$ . By Lemma 6.7 and the product rule, we have

$$-(d-1)^2(f_1^2 + f_2^2)^{\frac{5}{2}} \cdot \frac{\partial c(\mathbf{x})}{\partial x_i} = \frac{\partial \det H_0}{\partial x_i} \cdot (f_1^2 + f_2^2) - 3 \det H_0 \cdot (f_1 \cdot f_{1,i} + f_2 \cdot f_{2,i}).$$

This yields the stated polynomial equation.  $\square$

**Theorem 6.11** *The number of complex critical curvature points of a curve  $C$  defined by a general polynomial  $f$  of degree  $d$  is  $2d(3d-5)$ .*

**Proof** Recall that critical curvature points correspond to finite cusps of the evolute by Proposition 6.5. Piene, Riener and Shapiro prove in [148, Proposition 3.3] that, counting in the complex projective plane, the number of cusps on the evolute for a general plane curve  $C$  of degree  $d$  is  $6d^2 - 9d$ . As explained in the proof of Corollary 6.9, Salmon [161, §119] shows that  $d$  of these cusps lie at infinity. Therefore, the curve  $C$  has  $6d^2 - 9d - d = 2d(3d-5)$  complex critical curvature points.  $\square$

## 6.2 Algebraic Varieties

We study the curvature of smooth algebraic varieties in  $\mathbb{R}^n$  of any dimension. This will lead us to the notions of *second fundamental form* and *Weingarten map*. These are fundamental concepts in Riemannian geometry, and in standard textbooks they are presented in much more general contexts; see, for instance, [60, 123, 141, 147]. Thus, our main goal in this section is to formulate the second fundamental form and Weingarten map in terms of the polynomial equations that define the variety.

Let  $X \subset \mathbb{R}^n$  be a smooth algebraic variety of dimension  $m := \dim(X)$ . As in the case of plane curves, we consider a unit normal field  $N(\mathbf{x})$  for  $X$  and differentiate it along a tangent field  $T(\mathbf{x})$ . The main difference to the case of curves in the plane is that there can be several tangent and normal directions. Nevertheless, similar to (6.2), we define the curvature of  $X$  at a point  $\mathbf{x}$  in tangent direction  $T(\mathbf{x})$  and in normal direction  $N(\mathbf{x})$  to be  $\langle T(\mathbf{x}), T(\mathbf{x})_1 \cdot \frac{\partial N}{\partial x_1}(\mathbf{x}) + \dots + T(\mathbf{x})_n \cdot \frac{\partial N}{\partial x_n}(\mathbf{x}) \rangle$ . We will see that, as in the case of plane curves, this only depends on the values of  $T(\mathbf{x})$  and  $N(\mathbf{x})$  at a fixed point  $\mathbf{x}$ , but not on how those fields behave locally around  $\mathbf{x}$ . Let us work this out. Suppose that  $I(X) = \langle f_1, \dots, f_k \rangle$ . We denote the gradients of the defining polynomials  $f_i$  by  $\nabla f_i$  and their Hessians by  $H_i$  for  $i = 1, \dots, k$ . Let also  $J(\mathbf{x}) := (\nabla f_1 \dots \nabla f_k) \in \mathbb{R}^{n \times k}$  be the (transpose of) the Jacobian of the  $f_i$  at  $\mathbf{x}$ . A (local) smooth unit normal field on  $X$  is given by

$$N(\mathbf{x}) = \frac{J(\mathbf{x}) w(\mathbf{x})}{\|J(\mathbf{x}) w(\mathbf{x})\|} = \frac{1}{\|J(\mathbf{x}) w(\mathbf{x})\|} \sum_{i=1}^k w_i(\mathbf{x}) \cdot \nabla f_i(\mathbf{x}), \quad (6.6)$$

where  $w : X \rightarrow \mathbb{R}^k$  is smooth with  $w(\mathbf{x}) \notin \ker J(\mathbf{x})$  for all  $\mathbf{x}$ .

Differentiating (6.6) leads to

$$\frac{d}{d\mathbf{x}} N(\mathbf{x}) = \frac{1}{\|J(\mathbf{x}) w(\mathbf{x})\|} \sum_{i=1}^k w_i(\mathbf{x}) \cdot H_i(\mathbf{x}) + J(\mathbf{x}) R(\mathbf{x})$$

for some matrix valued function  $R(\mathbf{x})$ . Let us now fix a point  $\mathbf{x} \in X$  and denote the tangent vector by  $\mathbf{t} := T(\mathbf{x}) \in T_{\mathbf{x}}X$  and the normal vector by  $\mathbf{v} := J(\mathbf{x})w(\mathbf{x}) \in N_{\mathbf{x}}X$ . Since  $\mathbf{t}^\top J(\mathbf{x}) = 0$ , this implies

$$\left\langle \mathbf{t}, \mathbf{t}_1 \cdot \frac{\partial N}{\partial x_1} + \dots + \mathbf{t}_n \cdot \frac{\partial N}{\partial x_n} \right\rangle = \frac{1}{\|\mathbf{v}\|} \cdot \mathbf{t}^\top \left( \sum_{i=1}^k \mathbf{w}_i H_i \right) \mathbf{t}.$$

In particular, the right hand side only depends on the values of  $T(\mathbf{x})$  and  $N(\mathbf{x})$  at  $\mathbf{x}$ . This motivates the following definition.

**Definition 6.12** Using the notation above, we define the curvature of  $X$  at  $\mathbf{x}$  in tangent direction  $\mathbf{t} \in T_{\mathbf{x}}X$  and in normal direction  $\mathbf{v} \in N_{\mathbf{x}}X$  as

$$c(\mathbf{x}, \mathbf{t}, \mathbf{v}) := \frac{1}{\|\mathbf{v}\|} \mathbf{t}^\top \left( \sum_{i=1}^k \mathbf{w}_i \cdot H_i \right) \mathbf{t},$$

where  $\mathbf{v} = \sum_{i=1}^k \mathbf{w}_i \nabla f_i \in N_{\mathbf{x}}X$ .

In fact, for  $\mathbf{x} \in X$  and a fixed normal vector  $\mathbf{v} \in N_{\mathbf{x}}X$ , the curvature  $c(\mathbf{x}, \mathbf{t}, \mathbf{v})$  is a quadratic form on  $T_{\mathbf{x}}X$ . This quadratic form is called the *second fundamental form* of  $X$  at  $\mathbf{x}$  and  $\mathbf{v}$ . It is often denoted by

$$\Pi_{\mathbf{v}}(\mathbf{t}) := c(\mathbf{x}, \mathbf{t}, \mathbf{v}). \quad (6.7)$$

The associated linear map is called *Weingarten map*. We denote it by

$$L_{\mathbf{v}} : T_{\mathbf{x}}X \rightarrow T_{\mathbf{x}}X, \quad L_{\mathbf{v}}(\mathbf{t}) = P_{\mathbf{x}} \left( \sum_{i=1}^k \mathbf{w}_i \cdot H_i \cdot \mathbf{t} \right), \quad \text{where } \mathbf{v} = \sum_{i=1}^k \mathbf{w}_i \nabla f_i \quad (6.8)$$

and  $P_{\mathbf{x}} : \mathbb{R}^n \rightarrow T_{\mathbf{x}}X$  is the orthogonal projection onto the tangent space of  $X$  at  $\mathbf{x}$ . Since  $L_{\mathbf{v}}$  is a self-adjoint operator given by a real symmetric matrix, its eigenvalues are all real. If  $\|\mathbf{v}\| = 1$ , the eigenvalues of  $L_{\mathbf{v}}$  are called the *principal curvatures* of  $X$  at  $\mathbf{x}$  and in normal direction  $\mathbf{v}$ . The product of the principal curvatures is called the *Gauss curvature*; the arithmetic mean of the principal curvatures is called *mean curvature*. Since the principal curvatures are the critical points of the quadratic form  $\Pi_{\mathbf{v}}(\mathbf{t})$ , the *maximal curvature*

$$C(X) := \max_{\mathbf{x} \in X, \mathbf{t} \in T_{\mathbf{x}}X, \mathbf{v} \in N_{\mathbf{x}}X} c(\mathbf{x}, \mathbf{t}, \mathbf{v}) = \max_{\mathbf{x} \in X, \mathbf{v} \in N_{\mathbf{x}}X} \max_{\mathbf{t} \in T_{\mathbf{x}}X} c(\mathbf{x}, \mathbf{t}, \mathbf{v}) \quad (6.9)$$

is the maximum over all principal curvatures for varying  $(\mathbf{x}, \mathbf{v})$ .

**Example 6.13 (Hypersurfaces)** If  $X$  is defined by a single polynomial equation  $f(\mathbf{x}) = 0$ , we only have one normal direction (up to sign), and so the formula in Definition 6.12 can be written as

$$c(\mathbf{x}, \mathbf{t}) := \frac{1}{\|\nabla f(\mathbf{x})\|} \mathbf{t}^T H \mathbf{t},$$

where  $H$  is the Hessian of  $f$ . This generalizes the formula in Lemma 6.6.

Let us now focus on surfaces in  $\mathbb{R}^3$ . Let  $S \subset \mathbb{R}^3$  be a smooth algebraic surface and  $\mathbf{x} \in S$ . When the two principal curvatures of  $S$  at  $\mathbf{x}$  are equal, the point  $\mathbf{x}$  is called an *umbilic* or *umbilical point* of the surface  $S$ . Equivalently, the best second-order approximation of  $S$  at  $\mathbf{x}$  is a 2-sphere. Umbilical points can be formulated as the zeros of a system of polynomial equations, whose complex zeros are called *complex umbilics* of the surface  $S$ . Salmon [162] computed the number of complex umbilics of a general surface.

**Theorem 6.14** A general surface in  $\mathbb{R}^3$  defined by a polynomial of degree  $d$  has  $10d^3 - 28d^2 + 22d$  umbilics.

In the case of quadrics, we have results on the number of real umbilics and critical curvature points. Observe that rotations and translations do not affect the curvature, and that after a rotation and translation every quadric surface in  $\mathbb{R}^3$  has the form  $S = \{a_1x_1^2 + a_2x_2^2 + a_3x_3^2 = 1\}$ . The next theorem is proved in [27].

**Theorem 6.15** Consider a quadric surface  $S = \{a_1x_1^2 + a_2x_2^2 + a_3x_3^2 = 1\}$  with  $a_1, a_2, a_3 \neq 0$  and  $a_i \neq a_j$  for  $i \neq j$ . The number of real umbilics of  $S$  is

- 4 if  $S$  is an ellipsoid ( $a_1, a_2, a_3$  are positive) or a two-sheeted hyperboloid (one of the  $a_i$  is positive and two are negative);
- 0 if  $S$  is a one-sheeted hyperboloid (two of the  $a_i$  are positive and one is negative).

Similar to the case of plane curves, we call a point  $\mathbf{x} \in S$  a *critical curvature point* if one of the two principal curvatures of  $S$  attains a critical value at  $\mathbf{x}$ . The first observation is that umbilics are always critical curvature points. This was shown in [27]. The following theorems, also proved in [27], cover the case of quadrics.

**Theorem 6.16** A general quadric surface  $S \subset \mathbb{R}^3$  has 18 complex critical curvature points.

**Theorem 6.17** Consider a quadric surface  $S = \{a_1x_1^2 + a_2x_2^2 + a_3x_3^2 = 1\}$  with  $a_1, a_2, a_3 \neq 0$  and  $a_i \neq a_j$  for  $i \neq j$ . The number of real critical curvature points is

- 10 if  $S$  is an ellipsoid ( $a_1, a_2, a_3$  are positive);
- 4 if  $S$  is a one-sheeted hyperboloid (two of the  $a_i$  are positive and one is negative);
- 6 if  $S$  is a two-sheeted hyperboloid (one of the  $a_i$  is positive and two are negative).

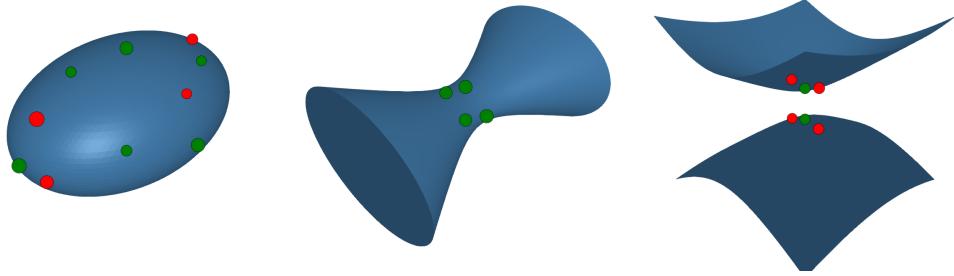


Fig. 6.5: The pictures illustrate Theorems 6.15 and 6.17. The figure on the left shows an ellipsoid with 4 red umbilics and 6 green critical curvature points. The umbilics are also critical curvature points, so that there are 10 in total. Similarly, the figure in the middle shows a one-sheeted hyperboloid with 4 green critical curvature points, and the figure on the left shows a two-sheeted hyperboloid with 4 red umbilics and 2 green critical curvature points (so 6 critical curvature points in total).

### 6.3 Volumes of Tubular Neighborhoods

We now turn to the problem of computing the volume of a *tubular neighborhood*. This is closely connected to curvature as we will see. We will discuss volumes of more general semialgebraic sets in Chapter 14. The tubular neighborhood of radius  $\varepsilon$  of a variety  $X \subset \mathbb{R}^n$  is

$$\text{Tube}(X, \varepsilon) := \{\mathbf{u} \in \mathbb{R}^n \mid d(\mathbf{u}, X) < \varepsilon\},$$

where  $d(\mathbf{u}, X) = \min_{\mathbf{x} \in X} \|\mathbf{u} - \mathbf{x}\|$  is the Euclidean distance from  $\mathbf{u}$  to  $X$ . There are several general formulas for upper bounds on the volume of  $\text{Tube}(X, \varepsilon)$  in the literature. For instance, Lotz [128] studied the case of a general complete intersection, and Bürgisser, Cucker and Lotz the case of a (possibly) singular hypersurface [37] of the sphere. The most general formula appeared in the paper by Basu and Lerario [13].

**Theorem 6.18** *Let  $X \subset \mathbb{R}^n$  be a real variety of dimension  $m$ . Let  $I(X) = \langle f_1, \dots, f_k \rangle$  be its ideal and let  $d := \max \deg(f_i)$ . Fix  $\mathbf{u} \in \mathbb{R}^n$  and  $r > 0$ . Denoting the ball of radius  $r$  around  $\mathbf{u}$  by  $B_r(\mathbf{u})$ , we have for every  $0 < \varepsilon \leq r/(4dm + m)$  that*

$$\frac{\text{vol}(\text{Tube}(X, \varepsilon) \cap B_r(\mathbf{u}))}{\text{vol}(B_r(\mathbf{u}))} \leq 4 \exp(1) \left( \frac{4nd\varepsilon}{r} \right)^{n-m}.$$

This theorem also holds for singular varieties. The proof of the theorem is based on approximating  $X$  in the Hausdorff topology by a sequence of smooth varieties  $(X_n)_{n \in \mathbb{N}}$  and showing that the volume of the tubular neighborhood of  $X_n$  can be controlled as  $n \rightarrow \infty$ . The volume of tubular neighborhoods of smooth varieties (in fact, of smooth submanifolds of  $\mathbb{R}^n$ ) is given by *Weyl's tube formula*. We derive this formula next. For a more detailed derivation and discussion, we refer to Weyl's original paper [180].

Let  $X \subset \mathbb{R}^n$  be smooth and  $\mathcal{N}X$  be the normal bundle of  $X$ . We denote the  $\varepsilon$ -normal bundle by

$$\mathcal{N}_\varepsilon X := \{(\mathbf{x}, \mathbf{v}) \in \mathcal{N}X \mid \|\mathbf{v}\| < \varepsilon\}.$$

Let us denote the *exponential map* by

$$\varphi_\varepsilon : \mathcal{N}_\varepsilon X \rightarrow \text{Tube}(X, \varepsilon), (\mathbf{x}, \mathbf{v}) \mapsto \mathbf{x} + \mathbf{v}. \quad (6.10)$$

**Definition 6.19** The *reach* of  $X$  or *injectivity radius* of  $X$  is defined as

$$\tau(X) := \sup\{\varepsilon > 0 \mid \varphi_\varepsilon \text{ is a diffeomorphism}\}. \quad (6.11)$$

For smooth  $X$ , the set in the right-hand side of (6.11) is non-empty and hence the reach  $\tau(X)$  is always positive; see e.g. [124, Theorem 6.24]. We will discuss further properties and applications of the reach in Chapters 7 and 15.

We can assume that  $X$  is compact (if not, we replace  $X$  by  $X \cap B$ , where  $B \subset \mathbb{R}^n$  is compact and full-dimensional). If  $\varepsilon < \tau(X)$ , the exponential map  $\varphi_\varepsilon$  is a diffeomorphism and we can pull back the integral over the tube to the normal bundle  $N_\varepsilon X$ . More specifically, let  $A$  be the matrix representation of the derivative of  $\varphi_\varepsilon$  with respect to orthonormal bases on the tangent space of a local trivialization of  $N_\varepsilon X$  around  $(\mathbf{x}, \mathbf{v})$  on the one side and  $\mathbb{R}^n$  on the other side. Then, we have

$$\text{vol}(\text{Tube}(X, \varepsilon)) = \int_{\text{Tube}(X, \varepsilon)} 1 \, d\mathbf{u} = \int_{\mathbf{x} \in X} \int_{\mathbf{v} \in N_\varepsilon X : \|\mathbf{v}\| < \varepsilon} |\det(A(\mathbf{x}, \mathbf{v}))| \, d\mathbf{v} \, d\mathbf{x}. \quad (6.12)$$

We compute the matrix  $A$ . The derivative of  $\varphi_\varepsilon$  is given by  $\frac{d\varphi_\varepsilon}{d(\mathbf{x}, \mathbf{v})}(\dot{\mathbf{x}}, \dot{\mathbf{v}}) = \dot{\mathbf{x}} + \dot{\mathbf{v}}$ . A local trivialization of  $N_\varepsilon X$  around  $(\mathbf{x}, \mathbf{v})$  has tangent space  $T_{\mathbf{x}}X \oplus N_{\mathbf{x}}X$ . The derivative of the normal component  $\dot{\mathbf{v}}$  splits into two components: one is due to movement in the normal space  $N_{\mathbf{x}}X$ , and one is due to curvature when moving in  $X$ . The derivative of  $\varphi_\varepsilon$  is thus given by the linear map  $T_{\mathbf{x}}X \oplus N_{\mathbf{x}}X \rightarrow \mathbb{R}^n$ ,  $(\mathbf{t}, \mathbf{z}) \mapsto \mathbf{t} + (\lambda \cdot L_w) \mathbf{t} + \mathbf{z}$ , where  $\mathbf{v} = \lambda \cdot \mathbf{w}$ ,  $\lambda = \|\mathbf{v}\|$ , and  $L_w$  is the Weingarten operator from (6.8). A matrix representation of the derivative of  $\varphi_\varepsilon$  with respect to orthonormal bases therefore is

$$A(\mathbf{x}, \mathbf{v}) = \begin{pmatrix} I_m + \lambda \cdot L_w & 0 \\ 0 & I_{n-m} \end{pmatrix}, \quad m = \dim X. \quad (6.13)$$

In fact, since  $\lambda = \|\mathbf{v}\| < \tau(X)$ , the eigenvalues of  $\lambda \cdot L_w$  are all of absolute value at most 1, so that  $|\det(A(\mathbf{x}, \mathbf{v}))| = \det(I_m + \lambda \cdot L_w)$ . The change of variables  $\mathbf{v} \rightarrow (\mathbf{w}, \lambda)$  has Jacobian determinant  $\lambda^{n-m-1}$ . Plugging all this into equation (6.12), we finally arrive at

$$\text{vol}(\text{Tube}(X, \varepsilon)) = \int_{\mathbf{x} \in X} \int_0^\varepsilon \int_{\mathbf{w} \in N_\varepsilon X : \|\mathbf{w}\|=1} \lambda^{n-m-1} \cdot \det(I_m + \lambda \cdot L_w) \, d\mathbf{w} \, d\lambda \, d\mathbf{x}.$$

Expanding the characteristic polynomial in this expression, we see that  $\text{vol}(\text{Tube}(X, \varepsilon))$  is given by a polynomial in  $\varepsilon$  of degree  $n$  whose coefficients are integrals of the principal minors of the Weingarten map  $L_w$  over  $X$ . Since  $L_{-\mathbf{v}} = -L_{\mathbf{v}}$ , the integrals over the odd-dimensional minors of  $L_{\mathbf{v}}$  vanish. All this leads to the following theorem.

**Theorem 6.20 (Weyl's tube formula [180])** *In the notation above, we have*

$$\text{vol}(\text{Tube}(X, \varepsilon)) = \sum_{0 \leq 2i \leq m} \kappa_{2i}(X) \cdot \varepsilon^{n-m+2i}.$$

The coefficients  $\kappa_{2i}(X)$  of this polynomial are called *curvature coefficients* of  $X$ . Explicitly, we have

$$\kappa_{2i}(X) = \frac{1}{n-m+2i} \int_{\mathbf{x} \in X} \int_{\mathbf{w} \in N_\varepsilon X : \|\mathbf{w}\|=1} m_{2i}(L_w) \, d\mathbf{w} \, d\mathbf{x},$$

where  $m_i(\cdot)$  denotes the sum of the  $2i$ -principal minors.

In fact, the curvature coefficient  $\kappa_0(X)$  is always equal to the  $m$ -dimensional volume of  $X$  times the volume of the unit ball  $B^{n-m} = \{\mathbf{x} \in \mathbb{R}^{n-m} \mid \|\mathbf{x}\| \leq 1\}$ . This leads to the following corollary.

**Corollary 6.21** *The volume of  $X$  can be obtained as the following limit:*

$$\text{vol}(X) = \lim_{\varepsilon \rightarrow 0} \frac{\text{vol}(\text{Tube}(X, \varepsilon))}{\varepsilon^{n-m} \cdot \text{vol}(B^{n-m})}.$$

We close this chapter by discussing Weyl's tube formula in two special cases.

**Example 6.22 (Curves)** If  $X \subset \mathbb{R}^n$  is a smooth algebraic curve, then

$$\text{vol}(\text{Tube}(X, \varepsilon)) = \varepsilon^{n-1} \cdot \text{vol}(B^{n-1}) \cdot \text{length}(X).$$

For instance, the volume of the  $\varepsilon$ -tube around a plane curve  $C$  is  $2\varepsilon \cdot \text{length}(C)$ .

**Example 6.23 (Surfaces in 3-space)** Let  $S = \{f(\mathbf{x}) = 0\} \subset \mathbb{R}^3$  be a compact smooth algebraic surface. Let  $\chi(S)$  be the Euler characteristic of  $S$ . Then, the volume of its  $\varepsilon$ -tube for  $\varepsilon < \tau(S)$  is

$$2\varepsilon \text{area}(S) + \varepsilon^3 \text{vol}(B^3) \chi(S).$$

Let us prove this. Weyl's tube formula implies that

$$\text{vol}(\text{Tube}(S, \varepsilon)) = 2\varepsilon \text{area}(S) + \varepsilon^3 \kappa_2(S).$$

The coefficient  $\kappa_2(S)$  is the integral of the Gauss curvature:

$$\kappa_2(S) = \frac{2}{3} \int_{\mathbf{x} \in S} \det(L_{N(\mathbf{x})}) \, d\mathbf{x},$$

where  $N(\mathbf{x}) = \|\nabla f(\mathbf{x})\|^{-1} \cdot \nabla f(\mathbf{x})$  denotes the normal field of  $S$  at  $\mathbf{x}$  given by the gradient of  $f$ . Notice that  $S$  is orientable – the orientation is given by the normal field  $N(\mathbf{x})$ . The Gauss-Bonnet theorem (see, e.g., [123, Theorem 9.3.]) implies that the integral over the Gauss curvature is  $2\pi \cdot \chi(S)$ . Moreover, the volume of the three-dimensional ball is  $\text{vol}(B^3) = 4\pi/3$ .

## **Chapter 7**

### **Medial Axis and Reach**

The *medial axis*  $\text{Med}(X)$  of a subset  $X \subset \mathbb{R}^n$  is the set of points  $\mathbf{u} \in \mathbb{R}^n$  such that there exist at least two points on  $X$  minimizing the distance to  $\mathbf{u}$ . In other words, the medial axis is the locus of points in  $\mathbb{R}^n$ , where the projection to  $X$  is not well-defined. In this section, we consider the case when  $X \subset \mathbb{R}^n$  is a real smooth variety.

In Definition 6.19 we have introduced the reach  $\tau(X)$  of  $X$  as the supremum over all  $\varepsilon$ , such that the exponential map of  $X$  restricted to normal vectors of length  $\varepsilon$  is a diffeomorphism. In particular, this means that all points  $\mathbf{u} \in \mathbb{R}^n$  that have distance less than  $\tau(X)$  to  $X$  have a unique closest point on  $X$ . Consequently, the distance between  $X$  and its medial axis must be at least  $\tau(X)$ . This shows that the reach is intimately linked with the medial axis. We make this connection more precise in the following proposition. In fact, the proposition can even be used as a definition for the reach of singular varieties. We emphasize that Theorem 7.6 below provides yet another equivalent formulation of the reach.

**Proposition 7.1** *The distance from a real variety  $X$  to its medial axis  $\text{Med}(X)$  is the reach  $\tau(X)$ .*

**Proof** We study the exponential map  $\varphi_\varepsilon : \mathcal{N}_\varepsilon X \rightarrow \text{Tube}(X, \varepsilon)$ ,  $(\mathbf{x}, \mathbf{v}) \mapsto \mathbf{x} + \mathbf{v}$  defined in (6.10). For all  $\varepsilon < \tau(X)$ , that map  $\varphi_\varepsilon$  is a diffeomorphism. In that case, the tubular neighborhood  $\text{Tube}(X, \varepsilon)$  cannot contain any point from the medial axis  $\text{Med}(X)$ , because otherwise  $\varphi_\varepsilon$  could not be injective. Hence, the distance from  $X$  to its medial axis  $\text{Med}(X)$  is at least  $\tau(X)$ .

Now we show the reverse inequality. For that, let  $\varepsilon > 0$  be less than the distance from  $X$  to its medial axis  $\text{Med}(X)$ . We consider a point  $\mathbf{u} \in \text{Tube}(X, \varepsilon)$ . In particular,  $\mathbf{u}$  is outside the Euclidean closure of  $\text{Med}(X)$ . Thus, in a Euclidean neighborhood  $U \subset \text{Tube}(X, \varepsilon)$  of  $\mathbf{u}$ , the points have a unique closest point on  $X$ . This defines a smooth map  $U \rightarrow X$ ,  $\mathbf{u} \mapsto \mathbf{x}(\mathbf{u})$ . The map  $\psi_{\mathbf{u}} : U \rightarrow \mathcal{N}_\varepsilon X$ ,  $\mathbf{u} \mapsto (\mathbf{x}(\mathbf{u}), \mathbf{u} - \mathbf{x}(\mathbf{u}))$  is then smooth with smooth inverse  $\psi_{\mathbf{u}}^{-1} = \varphi_\varepsilon|_{\psi_{\mathbf{u}}(U)}$ . Via a partition of unity of  $\text{Tube}(X, \varepsilon)$  (see e.g. [124, Chapter 2]), we can obtain a global smooth inverse of  $\varphi_\varepsilon$  from the local inverses  $\psi_{\mathbf{u}}$ . Hence,  $\varphi_\varepsilon$  is a diffeomorphism and  $\varepsilon < \tau(X)$ .  $\square$

Proposition 7.1 implies that the medial axis is semialgebraic. We call its complex Zariski closure

$$M_X := \overline{\text{Med}(X)},$$

the *algebraic medial axis* of  $X$ .

**Example 7.2** Consider the parabola  $X = V(x_2 - x_1^2)$ . We compute the algebraic medial axis of  $X$ . If  $\mathbf{x} = (x_1, x_2) \in X$  minimizes the distance to a point  $\mathbf{u} = (u_1, u_2) \in \mathbb{R}^2$ , we must have  $\langle \mathbf{x} - \mathbf{u}, \mathbf{t} \rangle = 0$ , where  $\mathbf{t}$  spans the tangent space  $T_{\mathbf{x}}X$ . We use Macaulay2 [79]:

```
R = QQ[x1, x2, y1, y2, u1, u2];
fx = x2 - x1^2; fy = y2 - y1^2;
Jx = matrix {{x1-u1, x2-u2}, {diff(x1, fx), diff(x2, fx)}};
Jy = matrix {{y1-u1, y2-u2}, {diff(y1, fy), diff(y2, fy)}};
distxu = (x1 - u1)^2 + (x2 - u2)^2;
distyu = (y1 - u1)^2 + (y2 - u2)^2;
I = ideal {fx, fy, det(Jx), det(Jy), distxu - distyu};
K = saturate(I, ideal {x1-y1, x2-y2});
eliminate({x1, x2, y1, y2}, K)
```

This returns the ideal  $\langle u_1 \rangle$ . Indeed, the medial axis is  $\text{Med}(X) = \{(0, u_2) \mid u_2 \geq \frac{1}{2}\}$ , so that the algebraic medial axis is  $M_X = \{u_1 = 0\}$  (see also Figure 7.4 below).  $\diamond$

As the medial axis, the reach of a real variety is an algebraic notion. More concretely, if  $X$  is smooth and defined by rational polynomials, then its reach  $\tau(X)$  is an algebraic number over  $\mathbb{Q}$ . This was shown in [97, Proposition 3.14]. An example comes next.

**Example 7.3** The reach of the parabola  $X$  in Example 7.2 is  $\tau(X) = \frac{1}{2}$ . It is obtained as the distance between  $(0, 0) \in X$  and  $(0, \frac{1}{2})$ , which is a point in the Euclidean closure of  $\text{Med}(X)$ .

## 7.1 Bottlenecks

We can characterize the reach of a smooth variety  $X \subset \mathbb{R}^n$  in terms of maximal curvature and bottlenecks. Curvature is discussed in details in the previous chapter. We now introduce bottlenecks. Let  $\mathbf{x}, \mathbf{y} \in X$  be two distinct points. If  $\mathbf{x} - \mathbf{y}$  is normal to  $T_{\mathbf{x}}X$  (i.e., for all  $\mathbf{t} \in T_{\mathbf{x}}X$  we have  $\langle \mathbf{x} + \mathbf{t}, \mathbf{x} - \mathbf{y} \rangle = 0$ ) and also normal to  $T_{\mathbf{y}}X$ , then we call  $(\mathbf{x}, \mathbf{y})$  a *bottleneck*. Complex solutions to the corresponding system of polynomial equations are called complex bottleneck. Di Rocco, Eklund and Weinstein [58] gave an algorithm for computing the number of complex bottlenecks of a variety in terms of polar classes. The following theorem is their result for planar curves.

**Theorem 7.4** *If  $X$  is a general curve in the plane of degree  $d$ , it has  $\frac{1}{2}(d^4 - 5d^2 + 4d)$  complex bottlenecks.*

**Example 7.5** We compute bottlenecks of the Trott curve  $144(x^4 + y^4) - 225(x^2 + y^2) + 350x^2y^2 + 81 = 0$  using `HomotopyContinuation.jl` [29].

```
@var x y u v
f = 144*(x^4 + y^4) - 225*(x^2 + y^2) + 350*x^2*y^2 + 81
g = subs(f, x=>u, y=>v)
df = differentiate(f, [x; y])
dg = differentiate(g, [u; v])
N = [x-u; y-v]
bottlenecks = solve([f; g; det([N df]); det([N dg])])
```

By Theorem 7.4 there are  $\frac{1}{2}(d^4 - 5d^2 + 4d) = 96$  complex bottlenecks. We find that 36 of them are real. They are marked in Figure 7.1. We note here that one point can appear in more than one bottleneck.  $\diamond$

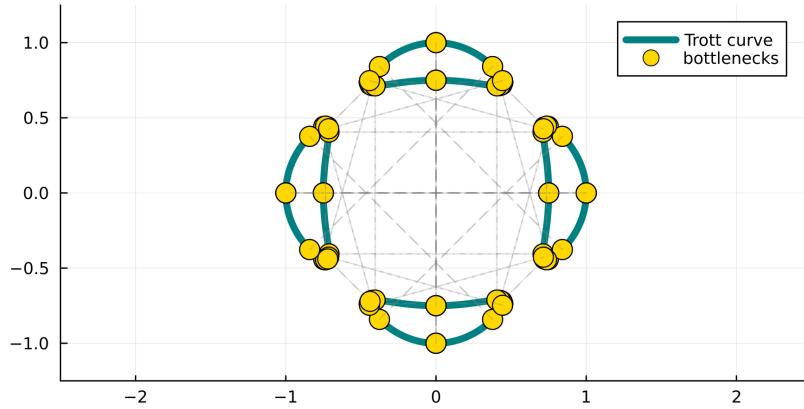


Fig. 7.1: Bottlenecks of the Trott curve  $144(x^4 + y^4) - 225(x^2 + y^2) + 350x^2y^2 + 81 = 0$  are displayed as grey normal lines with yellow endpoints.

The *width* of a bottleneck is  $b(\mathbf{x}, \mathbf{y}) := \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|$ . We denote the width of the smallest bottleneck of  $X$  by

$$B(X) := \min_{(\mathbf{x}, \mathbf{y}) \text{ bottleneck of } X} b(\mathbf{x}, \mathbf{y}).$$

The next theorem links the reach of a smooth variety  $X$  to its bottlenecks and its maximal curvature  $C(X)$  defined in (6.9).

**Theorem 7.6** *If  $X$  is smooth, then*

$$\tau(X) = \min \left\{ B(X), \frac{1}{C(X)} \right\}.$$

**Proof** Recall from (6.11) that the reach  $\tau(X)$  is the supremum over all  $\varepsilon > 0$  such that the exponential map  $\varphi_\varepsilon : N_\varepsilon X \rightarrow \text{Tube}(X, \varepsilon)$ ,  $(\mathbf{x}, \mathbf{v}) \mapsto \mathbf{x} + \mathbf{v}$  is a diffeomorphism. Let

$$\varepsilon := \tau(X). \quad (7.1)$$

Then, for every  $\varepsilon' > \varepsilon$ , the exponential map  $\varphi_{\varepsilon'}$  is not a diffeomorphism, which means that it is either not an immersion or it is not injective. Thus, there is a point  $\mathbf{u} = \mathbf{x} + \mathbf{v} \in \mathbb{R}^n$ , where  $(\mathbf{x}, \mathbf{v}) \in NX$ , at distance  $\varepsilon = \|\mathbf{v}\|$  from  $X$  such that, for every  $\varepsilon' > \varepsilon$ , either the derivative of  $\varphi_{\varepsilon'}$  at  $(\mathbf{x}, \mathbf{v})$  is not injective or  $\varphi_{\varepsilon'}^{-1}(\mathbf{u})$  has at least two elements.

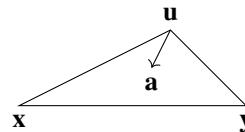
Suppose first that the derivative of each  $\varphi_{\varepsilon'}$  at  $(\mathbf{x}, \mathbf{v})$  is not injective. Then it follows from (6.13) that  $\varepsilon^{-1}$  is a principal curvature at  $\mathbf{x}$  in normal direction  $-\varepsilon^{-1}\mathbf{v}$ . Due to (7.1),  $\varepsilon$  is the smallest positive number with that property. Therefore, its inverse  $\varepsilon^{-1}$  must be the maximal curvature  $C(X)$ .

The remaining case to analyze is when each fiber  $\varphi_{\varepsilon'}^{-1}(\mathbf{u})$  is larger one and the derivative of  $\varphi_{\varepsilon'}$  at each point in that fiber is injective. We have two distinct points  $\mathbf{x}, \mathbf{y} \in X$  such that  $\mathbf{u} = \mathbf{x} + \mathbf{v} = \mathbf{y} + \mathbf{w}$ , where  $(\mathbf{v}, \mathbf{w}) \in NX$  and  $\delta := \|\mathbf{w}\| \leq \varepsilon$ . We distinguish two subcases. First, we assume that  $\mathbf{u}$  lies on the line spanned by  $\mathbf{x}$  and  $\mathbf{y}$ . Then,  $(\mathbf{x}, \mathbf{y})$  is a bottleneck. Moreover,  $\mathbf{u}$  must be the midpoint between  $\mathbf{x}$  and  $\mathbf{y}$ , since otherwise there would be a  $\sigma$  with  $b(\mathbf{x}, \mathbf{y}) < \sigma < \varepsilon$  and the fiber of the midpoint  $\frac{1}{2}(\mathbf{x} + \mathbf{y})$  under  $\varphi_\sigma$  would contain at least two points, but the latter implies  $\tau(X) \leq \sigma < \varepsilon$ ; a contradiction to (7.1). Hence,  $\varepsilon = b(\mathbf{x}, \mathbf{y})$  and due to (7.1) there cannot be any smaller bottleneck, so that  $\varepsilon = B(X)$  and we are done.

Second, we assume that  $\mathbf{x}, \mathbf{y}, \mathbf{u}$  form a triangle. Since the derivative of  $\varphi_{\varepsilon'}$  is injective at both  $(\mathbf{x}, \mathbf{v})$  and  $(\mathbf{y}, \mathbf{w})$ , the inverse function theorem implies that existence of two locally defined and smooth maps  $\mathbf{u} \mapsto \mathbf{x}(\mathbf{u})$  and  $\mathbf{u} \mapsto \mathbf{y}(\mathbf{u})$  that project locally around  $\mathbf{u}$  to  $X$ . These define two local smooth functions  $d_x(\mathbf{u}) := \|\mathbf{u} - \mathbf{x}(\mathbf{u})\|$  and  $d_y(\mathbf{u}) := \|\mathbf{u} - \mathbf{y}(\mathbf{u})\|$  that locally measure the distance to  $X$ . Their gradients are

$$\nabla d_x(\mathbf{u}) = \varepsilon^{-1}(\mathbf{u} - \mathbf{x}) \quad \text{and} \quad \nabla d_y(\mathbf{u}) = \delta^{-1}(\mathbf{u} - \mathbf{y});$$

see e.g. [78, Lemma 2.11] and also Remark 7.7. Let  $\mathbf{a}$  be a unit norm vector that starts at  $\mathbf{u}$ , points inside the triangle with vertices  $\mathbf{x}, \mathbf{y}, \mathbf{u}$  such that  $\langle \mathbf{u} - \mathbf{x}, \mathbf{a} \rangle < 0$  and  $\langle \mathbf{u} - \mathbf{y}, \mathbf{a} \rangle < 0$ :



We get that the partial derivatives of  $d_x(\mathbf{u})$  and  $d_y(\mathbf{u})$  in direction  $\mathbf{a}$  satisfy  $\frac{\partial d_x(\mathbf{u})}{\partial \mathbf{a}} < 0$  and  $\frac{\partial d_y(\mathbf{u})}{\partial \mathbf{a}} < 0$ . This means that, when we move from  $\mathbf{u}$  in direction  $\mathbf{a}$ , the local distances from  $\mathbf{u}$  to  $X$  both decrease. Consequently, there is a  $\sigma < \varepsilon$  such that  $\varphi_\sigma$  is not injective, hence  $\tau(X) \leq \sigma < \varepsilon$ . This contradicts (7.1) and so  $\mathbf{x}, \mathbf{y}, \mathbf{u}$  cannot form a triangle.  $\square$

**Remark 7.7** As in the proof of Theorem 7.6 let  $d$  be a (locally) defined projection map to a variety  $X$ , such that  $\mathbf{x} = d(\mathbf{u})$ . An informal argument that shows  $\nabla d(\mathbf{u}) = (\mathbf{u} - \mathbf{x})/\|\mathbf{u} - \mathbf{x}\|$  and that can be made rigorous goes as follows: Let us write  $\mathbf{v} := (\mathbf{u} - \mathbf{x})/\|\mathbf{u} - \mathbf{x}\|$  for the unit normal direction. If we move from  $\mathbf{u}$  infinitesimally in a direction that is perpendicular to  $\mathbf{v}$ , the distance  $d(\mathbf{u})$  does not change. This means

that the derivative of  $d(\mathbf{u})$  in a direction perpendicular to  $\mathbf{v}$  is zero, hence the gradient of  $d(\mathbf{u})$  must be a multiple of  $\mathbf{v}$ . On the other hand,  $d(\mathbf{u} + t \cdot \mathbf{v}) = d(\mathbf{u}) + t$ , which shows that the derivative of  $d(\mathbf{u})$  in direction  $\mathbf{v}$  is 1. Consequently,  $\nabla d(\mathbf{u}) = \mathbf{v}$ . In particular, the Weingarten map  $L_{\mathbf{v}}$  can be obtained via the second derivative of  $d(\mathbf{u})$ . This is worked out in (7.5) below.

## 7.2 Offset Hypersurfaces

This section is based on the article [97]. We write  $X_{\mathbb{C}}$  for the complex Zariski closure of  $X$ . We assume that  $X$  (and hence also  $X_{\mathbb{C}}$ ) is irreducible.

The *ED correspondence*  $\mathcal{E}_X$  of  $X$  is the Zariski closure of the set of tuples  $(\mathbf{x}, \mathbf{u})$  with  $\mathbf{x} \in X^{\text{sm}}$  such that  $\mathbf{x}$  is an ED-critical point for  $\mathbf{u}$ . We recall from Theorem 2.23 that

$$\mathcal{E}_X = \overline{\{(\mathbf{x}, \mathbf{x} + \mathbf{h}) \mid \mathbf{x} \in X^{\text{sm}}, (\mathbf{x}, \mathbf{h}) \in N_X\}} \subseteq X_{\mathbb{C}} \times \mathbb{C}^n.$$

The branch locus of the projection  $\mathcal{E}_X \rightarrow \mathbb{C}^n$  is called the *ED discriminant* or *evolute*. For planar curves, it coincides with the definition of the evolute above in Section 1.3. We denote it by  $\Sigma_X \subset \mathbb{C}^n$ .

For  $\varepsilon \in \mathbb{C}$  and  $\mathbf{u} \in \mathbb{C}^n$ , the  $\varepsilon$ -sphere around  $\mathbf{u}$  is the variety  $S(\mathbf{u}, \varepsilon) := V(\|\mathbf{x} - \mathbf{u}\|^2 - \varepsilon^2)$ .

**Definition 7.8** The *offset correspondence* of  $X$  is

$$\text{OC}_X = (\mathcal{E}_X \times \mathbb{C}) \cap \{(\mathbf{x}, \mathbf{u}, \varepsilon) \in \mathbb{C}^n \times \mathbb{C}^n \times \mathbb{C} \mid \mathbf{x} \in S(\mathbf{u}, \varepsilon)\}.$$

That is,  $\text{OC}_X$  is the complex Zariski closure of the set of tuples  $(\mathbf{x}, \mathbf{u}, \varepsilon)$  such that  $\mathbf{x}$  is an ED critical point for  $\mathbf{u}$ , and  $\varepsilon^2$  is the squared Euclidean distance (over  $\mathbb{R}$ ) between  $\mathbf{x}$  and  $\mathbf{u}$ .

We consider the two coordinate projections  $\pi_1 : \text{OC}_X \rightarrow X_{\mathbb{C}}$  and  $\pi_2 : \text{OC}_X \rightarrow \mathbb{C}^n \times \mathbb{C}$ . Clearly,  $\pi_1$  is dominant, i.e.,  $\overline{\pi_1(\text{OC}_X)} = X_{\mathbb{C}}$ . However, the other projection is not:

**Definition 7.9** We denote  $\text{Off}_X := \overline{\pi_2(\text{OC}_X)} \subset \mathbb{C}^n \times \mathbb{C}$  and call it the *offset hypersurface* of  $X$ .

The next lemma justifies the name.

**Lemma 7.10**  $\text{codim } \text{Off}_X = 1$ .

**Proof** The ED correspondence  $\mathcal{E}_X$  is the Zariski closure of a vector bundle of rank  $\text{codim}(X)$  over  $X^{\text{sm}}$ , which shows that  $\dim \mathcal{E}_X = n$ . Since  $X$  is irreducible,  $\mathcal{E}_X$  is irreducible. The offset correspondence  $\text{OC}_X$  is the intersection of  $\mathcal{E}_X \times \mathbb{C}$ , which is also irreducible, with a hypersurface. This implies  $\dim \text{OC}_X = n$ . Because the Euclidean Distance Degree of  $X$  is finite, the projection  $\pi_2$  has finite fibers generically, which implies  $\dim \text{Off}_X = n$ .  $\square$

**Remark 7.11** For a fixed radius  $r > 0$ , we can study the level set  $\text{Off}_{X,r} \subset \mathbb{C}^n$  that is the intersection of the offset hypersurface  $\text{Off}_X$  with the hypersurface  $\varepsilon = r$ . Figure 7.3 shows the level sets for  $r = 0.5$  and  $r = 1.25$ , respectively, of the offset surface of a parabola.

The boundary of the tubular neighborhood  $\text{Tube}(X, r)$  is always contained in the real locus of  $\text{Off}_{X,r}$ . In fact, the Euclidean boundary of  $\text{Tube}(X, r)$  is  $\partial \text{Tube}(X, r) = \{\mathbf{u} \in \mathbb{R}^n \mid d(\mathbf{u}, X) = r\}$ . Thus, for any  $\mathbf{u} \in \partial \text{Tube}(X, r)$ , there is a closest point  $\mathbf{x} \in X$  with  $\|\mathbf{x} - \mathbf{u}\| = r$ . In particular,  $\mathbf{x}$  is an ED critical point for  $\mathbf{u}$  and  $(\mathbf{x}, \mathbf{u}, r) \in \text{OC}_X$ , i.e.,  $\mathbf{u} \in \text{Off}_{X,r}$ .

For the parabola in Figure 7.3, the real locus of  $\text{Off}_{X,0.5}$  is equal to  $\partial \text{Tube}(X, 0.5)$ , while  $\partial \text{Tube}(X, 1.25)$  is strictly contained in the real part of  $\text{Off}_{X,1.25}$ . In this example, that change of behavior is governed by the reach  $\tau(X) = 0.5$ .

It follows from Lemma 7.10 that  $\text{Off}_X$  is the zero set of a polynomial that we denote by  $g_X(\mathbf{u}, \varepsilon)$ ; i.e.,

$$\text{Off}_X = V(g_X) \subset \mathbb{C}^n \times \mathbb{C}.$$

We call it the *offset polynomial*. It is also known as the *ED polynomial*. It is studied in detail in [142].

**Example 7.12** Consider the parabola  $X = V(x_2 - x_1^2)$ . We compute the offset polynomial of the parabola in Macaulay2 [79].

```
R = QQ[x1, x2, u1, u2, eps];
f = x2 - x1^2; d = (x1-u1)^2 + (x2-u2)^2 - eps^2;
J = matrix {{x1-u1, x2-u2}, {diff(x1, f), diff(x2, f)}};
OC = ideal {f, det(J), d};
O = eliminate({x1, x2}, OC)
g = (gens O)_0_0
```

This gives  $g_X(\mathbf{u}, \varepsilon) = g_0(\mathbf{u}) + g_1(\mathbf{u})\varepsilon^2 + g_2(\mathbf{u})\varepsilon^4 + g_3(\mathbf{u})\varepsilon^6$ , where

$$\begin{aligned} g_0(\mathbf{u}) &= (u_1^2 - u_2)^2(16u_1^2 + 16u_2^2 - 8u_2 + 1), & g_2(\mathbf{u}) &= 48u_1^2 + 16u_2^2 + 32u_2 - 8, \\ g_1(\mathbf{u}) &= -48u_1^4 - 32u_1^2u_2^2 + 8u_1^2u_2 - 32u_2^3 - 20u_1^2 - 8u_2^2 + 8u_2 - 1, & g_3(\mathbf{u}) &= -16. \end{aligned}$$

Figure 7.2 shows the offset surface  $g(\mathbf{u}, \varepsilon) = 0$ . ◊

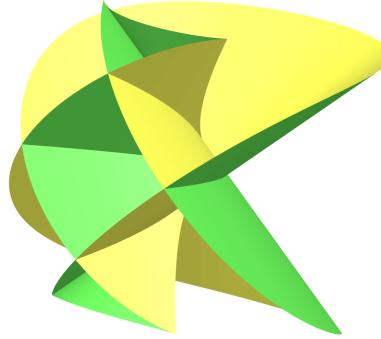


Fig. 7.2: The offset surface  $\text{Off}_X$  of the parabola. The surface is symmetric along the  $\varepsilon$ -axis, because only even powers of  $\varepsilon$  appear in the offset polynomial  $g(\mathbf{u}, \varepsilon)$ . The parabola itself is visible at level  $\varepsilon = 0$ . The image was created using [Surfer](#).

Let us study some properties of the offset polynomial.

**Proposition 7.13** 1. For a general  $\mathbf{u} \in \mathbb{C}^n$ , the zeros of  $g_X(\mathbf{u}, \varepsilon)$  are precisely  $\varepsilon = \pm\sqrt{\|\mathbf{u} - \mathbf{x}\|^2}$ , where  $\mathbf{x}$  ranges over all ED critical points for  $\mathbf{u}$ .  
2. The degree of the offset polynomial  $g_X(\mathbf{u}, \varepsilon)$  in  $\varepsilon$  is two times the Euclidean distance degree of  $X$ .

**Proof** First, we observe that the projection  $\text{Off}_X \rightarrow \mathbb{C}^n$ ,  $(\mathbf{u}, \varepsilon) \mapsto \mathbf{u}$  is dominant, because general points in  $\mathbb{C}^n$  have ED-critical points on  $X_{\mathbb{C}}$ . Take a general  $\mathbf{u} \in \mathbb{C}^n$ . Then,  $g_X(\mathbf{u}, \varepsilon) = 0$  if and only if there exists  $\mathbf{x} \in X_{\mathbb{C}}$  with  $(\mathbf{x}, \mathbf{u}) \in \mathcal{E}_X$  and  $\varepsilon^2 = \|\mathbf{x} - \mathbf{u}\|^2$ . Therefore,  $\varepsilon = \pm\sqrt{\|\mathbf{u} - \mathbf{x}\|^2}$ , where  $\mathbf{x}$  ranges over all ED critical points for  $\mathbf{u}$ . In particular, this shows that  $g_X(\mathbf{u}, \varepsilon)$  has  $2 \cdot \text{EDdeg}(X)$  many zeros for general  $\mathbf{u}$ .  $\square$

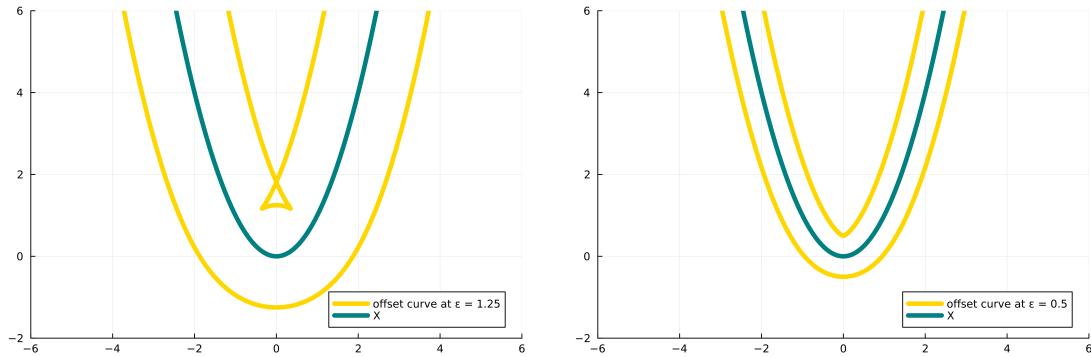


Fig. 7.3: The offset surface  $\text{Off}_X$  of the parabola intersected with the hypersurface  $\varepsilon = 1.25$  and  $\varepsilon = 0.5$ .

**Example 7.14** In Example 7.12, we see that  $g_X(\mathbf{u}, \varepsilon)$  has degree six in  $\varepsilon$ . This corresponds to the fact that the Euclidean distance degree of the parabola is three. For instance, any point  $\mathbf{u} \in \mathbb{R}^2$  in Figure 7.4 that is both above the green curve (the parabola) and the yellow curve (the evolute) has three real ED critical points on the parabola.  $\diamond$

The coefficients of the offset polynomial are studied in [142]. In particular, in that article one can find the following result.

**Theorem 7.15** *If the variety  $X$  is general enough and  $g_X(\mathbf{u}, \varepsilon) = c_0(\mathbf{u}) + c_1(\mathbf{u})\varepsilon + \dots + c_k(\mathbf{u})\varepsilon^k$  is its offset polynomial, then  $c_k(\mathbf{u})$  is constant.*

**Proof** See [142, Proposition 4.4].  $\square$

### 7.3 Offset Discriminant

We now understand that the offset polynomial  $g_X(\mathbf{u}, \varepsilon)$  encodes for a fixed  $\mathbf{u} \in \mathbb{R}^n$  the distances from  $\mathbf{u}$  to its ED-critical points on  $X$ . If  $\mathbf{u}$  is on the medial axis,  $g_X(\mathbf{u}, \varepsilon)$  must have a double root in  $\varepsilon$ . This motivates us to study the discriminant of the offset polynomial in  $\varepsilon$ .

**Definition 7.16** The *offset discriminant* is the polynomial

$$\delta_X(\mathbf{u}) := \text{Disc}_\varepsilon g_X(\mathbf{u}, \varepsilon).$$

Its zero set is denoted

$$\Delta_X^{\text{Off}} := V(\delta_X) \subset \mathbb{C}^n.$$

**Example 7.17** Using Macaulay2 [79] we compute the discriminant of the offset polynomial of the parabola from Example 7.12. We get

$$\delta_X(\mathbf{u}) = u_1^4 \cdot (u_1^2 - u_2) \cdot \delta_1(\mathbf{u})^6 \cdot \delta_2(\mathbf{u}),$$

where

$$\delta_1(\mathbf{u}) = -27u_1^2 + 2(2u_2 - 1)^3 \quad \text{and} \quad \delta_2(\mathbf{u}) = 16u_1^2 + (4u_2 - 1)^2.$$

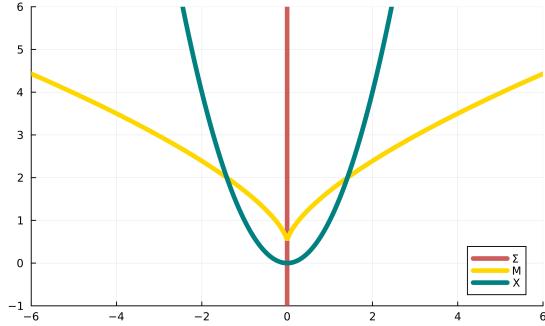


Fig. 7.4: The offset discriminant of the parabola  $X = V(x_2 - x_1^2)$  has three real components: the parabola itself (green), the algebraic medial axis  $M_X$  (the red vertical line) and the ED-discriminant or evolute  $\Sigma_X$  (the yellow cubic curve).

The factor  $\delta_2(\mathbf{u})$  has no real zeros, and  $u_1^2 - u_2$  is the polynomial of  $X$ . The real zero locus of the other factors  $u_1$  and  $\delta_1(\mathbf{u})$  are shown in Figure 7.4. The medial axis of the parabola is  $\text{Med}(X) = \{(0, u_2) \mid u_2 \geq \frac{1}{2}\}$  and  $u_1 = 0$  is the Zariski closure of  $\text{Med}(X)$ . The variety  $\delta_1(\mathbf{u}) = 0$  is the ED-discriminant or evolute of the parabola, also known as the *semicubical parabola*. Above the evolute a point  $\mathbf{u} \in \mathbb{R}^2$  has three real ED critical points on  $X$ , and below the evolute it has one real and two complex ED critical points.  $\diamond$

The discriminant  $\Delta_X^{\text{Off}}$  in the previous example we observed had three real components: the variety  $X$ , its algebraic medial axis and its evolute. We show that this is a general fact, following [97]. For that, we define the *bisector hypersurface*  $\text{Bis}_X$  of  $X$ : Writing  $\text{Bl}_X \subset \mathbb{C}^n \times \mathbb{C}$  for the branch locus of the projection  $\pi_2 : \text{OC}_X \rightarrow \mathbb{C}^n \times \mathbb{C}$ , the bisector hypersurface is the union of the branch points  $u$  when varying over all  $\varepsilon$ :

$$\text{Bis}_X := \bigcup_{(u, \varepsilon) \in \text{Bl}_X} u.$$

**Theorem 7.18** *The components of the offset discriminant are*

$$\Delta_X^{\text{Off}} = \text{Bis}_X \cup \Sigma_X \supseteq X_{\mathbb{C}} \cup M_X \cup \Sigma_X.$$

*Its real components are the real parts of  $X$ ,  $M_X$ , and  $\Sigma_X$ .*

**Proof** By Proposition 7.13, the offset discriminant is the locus of those  $\mathbf{u}$  such that  $g_X(\mathbf{u}, \varepsilon)$  has less than  $2 \times \text{EDdeg}(X)$  distinct complex solutions. This can happen due to two reasons:  $\mathbf{u}$  has either less than  $\text{EDdeg}(X)$  distinct ED critical points on  $X$  or two distinct ED critical points

$$\mathbf{x}_1 \neq \mathbf{x}_2 \text{ with } \|\mathbf{x}_1 - \mathbf{u}\|^2 = \|\mathbf{x}_2 - \mathbf{u}\|^2. \quad (7.2)$$

The first case is the ED discriminant  $\Sigma_X$  and second case the bisector hypersurface  $\text{Bis}_X$ . By definition, the medial axis  $\text{Med}(X)$  is contained in  $\text{Bis}_X$ , and thus we also have the inclusion  $M_X \subseteq \text{Bis}_X$ . Since the  $\varepsilon$  that come from zeros of  $g_X(\mathbf{u}, \varepsilon)$  come in signed pairs (cf. Proposition 7.13), we see that  $X \times \{0\}$  is doubly covered by the projection  $\pi_2$ , meaning that  $X \subseteq \text{Bis}_X$ . All other components of  $\text{Bis}_X$  besides  $X \cup M_X$  consist of non-real points  $\mathbf{u}$  that have ED critical points as in (7.2).  $\square$

Suppose now that  $\mathbf{u} \in \mathbb{R}^n \setminus \Delta_X^{\text{Off}}$  is a real point outside the offset discriminant. Using Remark 7.7 we can compute a unit normal field from the ED-polynomial. In fact, if  $(\mathbf{u}, \varepsilon) \in \mathbb{R}^n \times \mathbb{R}$  is a real zero of  $g_X$ ,

then it follows from Remark 7.7 that the gradient  $\frac{d\varepsilon}{du}(\mathbf{u}, \varepsilon)$  is a unit normal vector pointing from a real ED-critical point on  $X$  towards  $\mathbf{u}$ . Differentiating the equation  $g_X(\mathbf{u}, \varepsilon) = 0$ , we obtain

$$\nabla_\varepsilon := \frac{d\varepsilon}{du}(\mathbf{u}, \varepsilon) = -\left(\frac{\partial g_X}{\partial \varepsilon}\right)^{-1} \cdot \frac{\partial g_X}{\partial \mathbf{u}} \in \mathbb{R}^n. \quad (7.3)$$

Furthermore, differentiating (7.3) we obtain the following formula for the Hessian matrix of  $\varepsilon$ :

$$\frac{d^2\varepsilon}{du^2} = -\left(\frac{\partial g_X}{\partial \varepsilon}\right)^{-1} \cdot \left(\frac{\partial^2 g_X}{\partial \mathbf{u}^2} + \frac{\partial^2 g_X}{\partial \varepsilon^2} \cdot \nabla_\varepsilon (\nabla_\varepsilon)^T + 2\nabla_\varepsilon \left(\frac{\partial(g_X/\partial \varepsilon)}{\partial \mathbf{u}}\right)^T\right) \in \mathbb{R}^{n \times n}. \quad (7.4)$$

In fact, both the gradient and the Hessian of  $g_X$  are rational functions, whose denominator vanishes when  $\frac{\partial g_X}{\partial \varepsilon}(\mathbf{u}, \varepsilon) = 0$ . This is why we assume that  $\mathbf{u} \notin \Delta_X^{\text{Off}}$ . The next theorem shows that we can compute the second fundamental form of  $X$  from the Hessian of  $g_X$ .

**Theorem 7.19** *Let  $\mathbf{u} \in \mathbb{R}^n \setminus \Delta_X^{\text{Off}}$  and  $\varepsilon > 0$  with  $g_X(\mathbf{u}, \varepsilon) = 0$ . Let  $\mathbf{x} \in X$  be the (uniquely determined) ED-critical point corresponding to  $(\mathbf{u}, \varepsilon)$ .*

1.  $\frac{d\varepsilon}{du}(\mathbf{u}, \varepsilon)$  is a unit normal vector at  $\mathbf{x}$  pointing towards  $\mathbf{u}$ .

2. Let  $\mathbf{t} \in T_{\mathbf{x}}X$  be a tangent vector. Then, the second fundamental form of  $X$  at  $\mathbf{x}$  evaluated at  $\mathbf{t}$  is

$$\Pi_{\mathbf{u}-\mathbf{x}}(\mathbf{t}) = \lim_{\substack{s \rightarrow 0 \\ s > 0}} \mathbf{t}^T \left( \frac{d^2\varepsilon}{du^2}(\mathbf{x} + s(\mathbf{u} - \mathbf{x}), s\varepsilon) \right) \mathbf{t}.$$

**Proof** The first item follows from Remark 7.7 and we have discussed this above. We need to prove the second item.

Since  $\mathbf{u} \notin \Delta_X^{\text{Off}}$ , there can only be finitely many points on the line through  $\mathbf{x}$  and  $\mathbf{u}$  that belong to  $\Delta_X^{\text{Off}}$ . This implies that there exists  $s' > 0$ , such that  $\mathbf{x} + s(\mathbf{u} - \mathbf{x}) \notin \Delta_X^{\text{Off}}$  for all  $0 < s < s'$ . The first item of the theorem implies that for all for all  $0 < s < s'$  we have  $\frac{d\varepsilon}{du}(\mathbf{x} + s(\mathbf{u} - \mathbf{x}), s\varepsilon)$  is a unit normal vector at  $\mathbf{x}$  pointing towards  $\mathbf{u}$ .

Fix now  $0 < s < s'$ . By definition of the second fundamental form (6.7) we have that  $\Pi_{\mathbf{u}-\mathbf{x}}(\mathbf{t})$  is the directional derivative of the unit normal field above in direction  $\mathbf{t}$ . That is, we have

$$\Pi_{\mathbf{u}-\mathbf{x}}(\mathbf{t}) = \mathbf{t}^T \left( \frac{d}{dx} \frac{d\varepsilon}{du}(\mathbf{x} + s(\mathbf{u} - \mathbf{x}), s\varepsilon) \right) \mathbf{t} \quad \text{for all } 0 < s < s'. \quad (7.5)$$

We have shown in (6.13) that

$$\frac{d\mathbf{u}}{dx} = I_n + \varepsilon \cdot \begin{pmatrix} L_v & 0 \\ 0 & 0 \end{pmatrix},$$

where  $L_v$  is the Weingarten map of  $X$  at  $\mathbf{x}$  in normal direction  $\mathbf{v} = (\mathbf{u} - \mathbf{x})/\|\mathbf{u} - \mathbf{x}\|$ . Hence, applying the chain rule yields

$$\frac{d}{dx} \frac{d\varepsilon}{du} = \frac{d^2\varepsilon}{du^2} \cdot \frac{d\mathbf{u}}{dx} = \frac{d^2\varepsilon}{du^2} \left( I_n + \varepsilon \cdot \begin{pmatrix} L_v & 0 \\ 0 & 0 \end{pmatrix} \right).$$

Evaluating this equation at  $(\mathbf{x} + s(\mathbf{u} - \mathbf{x}), s\varepsilon)$ , we obtain

$$\frac{d^2\varepsilon}{du^2}(\mathbf{x} + s(\mathbf{u} - \mathbf{x}), s\varepsilon) = \frac{d}{dx} \frac{d\varepsilon}{du}(\mathbf{x} + s(\mathbf{u} - \mathbf{x}), s\varepsilon) - s\varepsilon \cdot \left( \frac{d^2\varepsilon}{du^2} \cdot \begin{pmatrix} L_{\nabla\varepsilon} & 0 \\ 0 & 0 \end{pmatrix} \right) (\mathbf{x} + s(\mathbf{u} - \mathbf{x}), s\varepsilon).$$

If we multiply both sides of this equation from the left by  $\mathbf{t}^T$  and from the right by  $\mathbf{t}$  and then take the limit  $s \rightarrow 0$ , the formula (7.5) implies that we obtain  $\Pi_{\mathbf{u}-\mathbf{x}}(\mathbf{t})$  on the right-hand side. This yields the stated formula.  $\square$

**Example 7.20** We compute the expression (7.3) for the parabola  $X = V(x_2 - x_1^2)$  from Example 7.12. Namely,  $\frac{d\varepsilon}{du}(\mathbf{u}, \varepsilon) = (h_1, h_2)$ , where

$$h_1 = \frac{-96u_1\varepsilon^4 + (192u_1^3 + 64u_1u_2^2 - 16u_1u_2 + 40u_1)\varepsilon^2 - 4u_1(u_1^2 - u_2)(24u_1^2 + 16u_2^2 - 16u_2 + 1)}{-96\varepsilon^5 + (192u_1^2 + 64u_2^2 + 128u_2 - 32)\varepsilon^3 + (-96u_1^4 - 64u_1^2u_2^2 + 16u_1^2u_2 - 64u_2^3 - 40u_1^2 - 16u_2^2 + 16u_2 - 2)\varepsilon}$$

$$h_2 = \frac{(-32u_2 - 32)\varepsilon^4 + (64u_1^2u_2 - 8u_1^2 + 96u_2^2 + 16u_2 - 8)\varepsilon^2 - 2(u_1^2 - u_2)(16u_1^2u_2 - 20u_1^2 - 32u_2^2 + 12u_2 - 1)}{-96\varepsilon^5 + (192u_1^2 + 64u_2^2 + 128u_2 - 32)\varepsilon^3 + (-96u_1^4 - 64u_1^2u_2^2 + 16u_1^2u_2 - 64u_2^3 - 40u_1^2 - 16u_2^2 + 16u_2 - 2)\varepsilon}.$$

For instance, if we plug in  $(u_1, u_2, \varepsilon) = (0, \frac{1}{4}, \frac{1}{4})$ , we obtain  $(h_1, h_2) = (0, 1)$ , which is the unit normal vector on the parabola at  $\mathbf{x} = (0, 0)$  pointing towards  $\mathbf{u} = (0, \frac{1}{4})$ . Now we compute the second derivative of  $\varepsilon$  using the formula in (7.4), evaluate it at  $(u_1, u_2, \varepsilon) = (0, s, s)$  and let  $s \rightarrow 0, s > 0$ . This yields the matrix

$$A := \begin{pmatrix} -2 & 0 \\ 0 & 0 \end{pmatrix}.$$

By (7.5), the (signed) curvature of the parabola at  $\mathbf{x} = (0, 0)$  is  $\mathbf{t}^T A \mathbf{t} = -2$ , where  $\mathbf{t} = (1, 0)^T$  is the tangent direction of  $X$  at  $\mathbf{x}$  (the negative sign is because the normal field points “inwards” which causes the derivative of the normal field to point in the opposite direction; this can be seen in Figure 6.2). We remark that in Example 7.3 we have shown that the reach of  $X$  is  $\frac{1}{2}$ . This confirms Theorem 7.6, which states that the reach of the parabola is equal to the inverse of its maximal curvature  $C(X)$  (a parabola has no real bottlenecks). Indeed, the curvature of a parabola is maximal at the apex.  $\diamond$

## **Chapter 8**

### **Voronoi Cells**

Every real algebraic variety determines a Voronoi decomposition of its ambient Euclidean space. Each Voronoi cell is a convex semialgebraic set in the normal space of the variety at a point. In this chapter we study such Voronoi cells of algebraic varieties, with primary focus on their algebraic boundaries.

Metric algebraic geometry is concerned with properties of real algebraic varieties that depend on a distance metric. Key concepts include the Euclidean distance degree [62], distance function [143], bottlenecks [59, 70], reach, offset hypersurfaces, medial axis [96], and cut locus [51]. Voronoi cells are also an important topic in metric algebraic geometry. We here consider them only for the Euclidean metric, but it also makes much sense to study Voronoi cells with respect to Kullback-Leibler divergence [5] or Wasserstein distance [16].

## 8.1 Voronoi Basics

We begin with the familiar case when the given variety  $X$  is a finite subset of the Euclidean space  $\mathbb{R}^n$ . The *Voronoi cell* of a point  $\mathbf{y} \in X$  consists of all points whose closest point in  $X$  is  $\mathbf{y}$ , i.e.

$$\text{Vor}_X(\mathbf{y}) := \{ \mathbf{u} \in \mathbb{R}^n : \mathbf{y} \in \arg \min_{\mathbf{x} \in X} \|\mathbf{x} - \mathbf{u}\|^2 \}. \quad (8.1)$$

This is a convex polyhedron with at most  $|X| - 1$  facets. The study of these cells, and how they depend on the configuration  $X$ , is ubiquitous in computational geometry and its numerous applications.

**Proposition 8.1** *The Voronoi cell of a point  $\mathbf{y}$  in the finite set  $X \subset \mathbb{R}^n$  is the polyhedron*

$$\text{Vor}_X(\mathbf{y}) = \{ \mathbf{u} \in \mathbb{R}^n : \mathbf{u} \cdot (\mathbf{x} - \mathbf{y}) \leq \frac{1}{2} (\|\mathbf{x}\|^2 - \|\mathbf{y}\|^2) \text{ for all } \mathbf{x} \in X \setminus \{\mathbf{y}\} \}. \quad (8.2)$$

**Proof** By definition,  $\text{Vor}_X(\mathbf{y})$  consists of all points  $\mathbf{u}$  such that  $\|\mathbf{x} - \mathbf{u}\|^2 - \|\mathbf{y} - \mathbf{u}\|^2$  is nonnegative for all  $\mathbf{x} \in X \setminus \{\mathbf{y}\}$ . But, this expression is equal to  $\|\mathbf{x}\|^2 - \|\mathbf{y}\|^2 - 2\mathbf{u} \cdot (\mathbf{x} - \mathbf{y})$ . The main point is that the quadratic term drops out, so the expression is linear in  $\mathbf{u}$ .  $\square$

The collection of all Voronoi cells, as  $\mathbf{y}$  ranges over the set  $X$ , is known as the *Voronoi diagram* of  $X$ . The Voronoi diagram is a polyhedral subdivision of  $\mathbb{R}^n$  into finitely many convex cells.

We now shift gears, and we replace the finite set  $X$  by a real algebraic variety of positive dimension. As before, the ambient space is  $\mathbb{R}^n$  with its standard Euclidean metric. We seek the Voronoi diagram  $\{\text{Vor}_X(\mathbf{y})\}_{\mathbf{y} \in X}$  in  $\mathbb{R}^n$  where  $\mathbf{y}$  runs over all (infinitely many) points in  $X$ .

One approach is to take a large but finite sample from  $X$  and to consider the Voronoi diagram of that sample. This is a finite approximation to the desired limit object. By taking finer and finer samples, the Voronoi diagram should converge nicely to a subdivision with infinitely many regions. The Voronoi cells in the limit are convex sets. However, for  $n \geq 3$ , they are generally not polyhedra. This process was studied by Brandt and Weinstein in [25] for the case when  $n = 2$  and  $X$  is a curve. In [25, Figure 1] we see this for a quartic curve. The authors posted a delightful YouTube video, called *Mathemaddies' Ice Cream Map*. Please do watch that movie! Their curve  $X$  is the shoreline that separates the city of Berkeley from the San Francisco Bay. One hopes to find many ice cream shops at the shore.

Let  $X$  be a real algebraic variety of codimension  $c$  in  $\mathbb{R}^n$ , and consider a point  $\mathbf{y} \in X$ . The Voronoi cell  $\text{Vor}_X(\mathbf{y})$  is defined as before. It consists of all points  $\mathbf{u}$  in  $\mathbb{R}^n$  such that  $\mathbf{y}$  is closer or equal to  $\mathbf{u}$  than any other point  $\mathbf{x} \in X$ . The equation (8.2) still holds, and we conclude that  $\text{Vor}_X(\mathbf{y})$  is a convex set.

**Proposition 8.2** *Suppose that  $\mathbf{y}$  is a smooth point of the variety  $X$ . Then its Voronoi cell  $\text{Vor}_X(\mathbf{y})$  is a convex semialgebraic set of dimension  $c$ . This Voronoi cell is contained in the normal space*

$$N_X(\mathbf{y}) = \{ \mathbf{u} \in \mathbb{R}^n : \mathbf{u} - \mathbf{y} \text{ is perpendicular to the tangent space of } X \text{ at } \mathbf{y} \} \simeq \mathbb{R}^c.$$

**Proof** Fix  $\mathbf{u} \in \text{Vor}_X(\mathbf{y})$ . Consider any point  $\mathbf{x}$  in  $X$  that is close to  $\mathbf{y}$ , and set  $\mathbf{v} = \mathbf{x} - \mathbf{y}$ . The inequality in (8.2) implies  $\mathbf{u} \cdot \mathbf{v} \leq \frac{1}{2}(||\mathbf{y} + \mathbf{v}||^2 - ||\mathbf{y}||^2) = \mathbf{y} \cdot \mathbf{v} + \frac{1}{2}||\mathbf{v}||^2$ . For any  $\mathbf{w}$  in the tangent space of  $X$  at  $\mathbf{y}$ , there exists  $\mathbf{v} = \epsilon\mathbf{w} + O(\epsilon^2)$  such that  $\mathbf{x} = \mathbf{y} + \mathbf{v}$  is in  $X$ . The inequality above yields  $\mathbf{u} \cdot \mathbf{w} \leq \mathbf{y} \cdot \mathbf{w}$ , and the same with  $-\mathbf{w}$  instead of  $\mathbf{w}$ . Then  $(\mathbf{u} - \mathbf{y}) \cdot \mathbf{w} = 0$ , and hence  $\mathbf{u}$  is in the normal space  $N_X(\mathbf{y})$ . We already argued that  $\text{Vor}_X(\mathbf{y})$  is convex. It is semialgebraic, by Tarski's Theorem on Quantifier Elimination. This allows us to eliminate  $\mathbf{x}$  from the formula (8.2). Finally, the Voronoi cell  $\text{Vor}_X(\mathbf{y})$  is full-dimensional in the  $c$ -dimensional space  $N_X(\mathbf{y})$  because every point  $\mathbf{u}$  in an  $\epsilon$ -neighborhood of  $\mathbf{y}$  has a unique closest point in  $X$ . Moreover, if  $\mathbf{u} \in N_X(\mathbf{y})$  then that closest point must be  $\mathbf{y}$ , by the same inequality as above.  $\square$

The topological boundary of the Voronoi cell  $\text{Vor}_X(\mathbf{y})$  in the normal space  $N_X(\mathbf{y})$  is denoted by  $\partial\text{Vor}_X(\mathbf{y})$ . It consists of all points in  $N_X(\mathbf{y})$  that have at least two closest points in  $X$ , including  $\mathbf{y}$ . We are interested in the *algebraic boundary*  $\partial_{\text{alg}}\text{Vor}_X(\mathbf{y})$ . This is the hypersurface in the complex affine space  $N_X(\mathbf{y})_{\mathbb{C}} \simeq \mathbb{C}^c$  obtained as the Zariski closure of  $\partial\text{Vor}_X(\mathbf{y})$  over the field of definition of  $X$ . The degree of this hypersurface is denoted  $\delta_X(\mathbf{y})$  and called the *Voronoi degree* of  $X$  at  $\mathbf{y}$ . If  $X$  is irreducible and  $\mathbf{y}$  is a general point on  $X$ , then this degree does not depend on the choice of  $\mathbf{y}$ .

**Example 8.3 (Surfaces in 3-space)** Fix a general polynomial  $f \in \mathbb{Q}[x_1, x_2, x_3]$  of degree  $d \geq 2$  and let  $X = V(f)$  be its surface in  $\mathbb{R}^3$ . The normal space at a general point  $\mathbf{y} \in X$  is the line  $N_X(\mathbf{y}) = \{\mathbf{y} + \lambda(\nabla f)(\mathbf{y}) : \lambda \in \mathbb{R}\}$ . The Voronoi cell  $\text{Vor}_X(\mathbf{y})$  is a (possibly unbounded) line segment in  $N_X(\mathbf{y})$  that contains  $\mathbf{y}$ . The boundary  $\partial\text{Vor}_X(\mathbf{y})$  consists of at most two points from among the zeros of an irreducible polynomial in  $\mathbb{Q}[\lambda]$ . We shall see that this univariate polynomial has degree  $d^3 + d - 7$ . Its complex zeros form the algebraic boundary  $\partial_{\text{alg}}\text{Vor}_X(\mathbf{y})$ . Thus, the Voronoi degree of the surface  $X$  is  $d^3 + d - 7$ .

Note that, in this example, our hypothesis “over the field of definition” becomes important. The  $\mathbb{Q}$ -Zariski closure of one boundary point is the collection of all  $d^3 + d - 7$  points in  $\partial_{\text{alg}}\text{Vor}_X(\mathbf{y})$ .

For a numerical example, we take the degree to be  $d = 2$  and we fix  $\mathbf{y} = (0, 0, 0)$ . Our quadric is  $f = x_1^2 + x_2^2 + x_3^2 - 3x_1x_2 - 5x_1x_3 - 7x_2x_3 + x_1 + x_2 + x_3$ . Let  $r_0 \approx -0.209$ ,  $r_1 \approx -0.107$ ,  $r_2 \approx 0.122$  be the roots of the cubic polynomial  $368\lambda^3 + 71\lambda^2 - 6\lambda - 1$ . The Voronoi cell  $\text{Vor}_X(\mathbf{y})$  is the line segment connecting the points  $(r_1, r_1, r_1)$  and  $(r_2, r_2, r_2)$ . The topological boundary  $\partial\text{Vor}_X(\mathbf{y})$  consists of the two points  $(r_1, r_1, r_1)$  and  $(r_2, r_2, r_2)$ , whereas the algebraic boundary  $\partial_{\text{alg}}\text{Vor}_X(\mathbf{y})$  also contains  $(r_0, r_0, r_0)$ .

The cubic polynomial in the unknown  $\lambda$  was found with the algebraic method that is described in the next section. Namely, the Voronoi ideal in (8.3) equals  $\text{Vor}_I(0) = \langle u_1 - u_3, u_2 - u_3, 368u_3^3 + 71u_3^2 - 6u_3 - 1 \rangle$ . This is a maximal ideal in  $\mathbb{Q}[u_1, u_2, u_3]$ , and it defines a field extension of degree 3 over  $\mathbb{Q}$ .

**Example 8.4 (Curves in 3-space)** Let  $X$  be a general algebraic curve in  $\mathbb{R}^3$ . For  $\mathbf{y} \in X$ , the Voronoi cell  $\text{Vor}_X(\mathbf{y})$  is a convex set in the normal plane  $N_X(\mathbf{y}) \simeq \mathbb{R}^2$ . Its algebraic boundary  $\partial_{\text{alg}}\text{Vor}_X(\mathbf{y})$  is a plane curve of degree  $\delta_X(\mathbf{y})$ . This Voronoi degree can be expressed in terms of the degree and genus of  $X$ . Specifically, this degree is 12 when  $X$  is the intersection of two general quadrics in  $\mathbb{R}^3$ . Figure 8.1 shows one such quartic space curve  $X$  together with the normal plane at a point  $\mathbf{y} \in X$ . The Voronoi cell  $\text{Vor}_X(\mathbf{y})$  is the planar convex region highlighted on the right. Its algebraic boundary  $\partial_{\text{alg}}\text{Vor}_X(\mathbf{y})$  is a curve of degree  $\delta_X(\mathbf{y}) = 12$ . The topological boundary  $\partial\text{Vor}_X(\mathbf{y})$  is only a very small subset of that algebraic boundary.

## 8.2 Computing Algebraic Boundaries

We study the Voronoi decomposition to answer the question for any point in ambient space, “What point on the variety  $X$  am I closest to?” Another question one might ask is, “How far do we have to get away from  $X$  before there is more than one answer to the closest point question?” The union of the boundaries of the Voronoi cells is the locus of points in  $\mathbb{R}^n$  that have more than one closest point on  $X$ . This set is called the *medial axis* (or *cut locus*) of the variety.

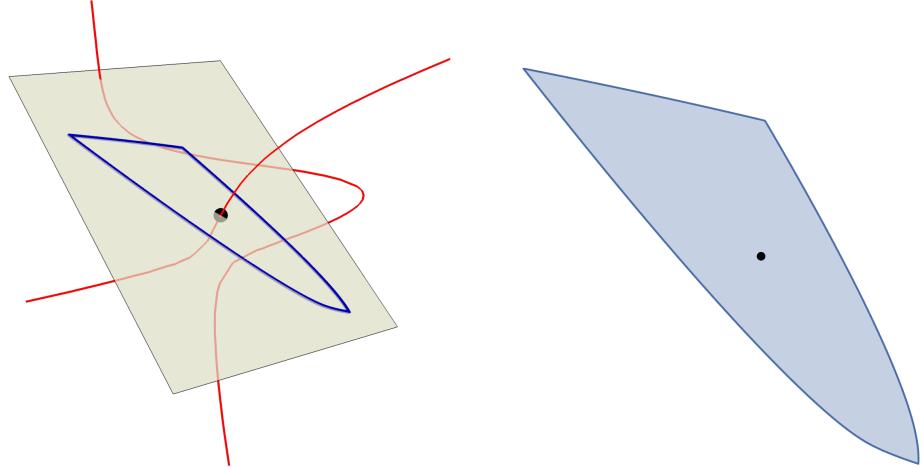


Fig. 8.1: A quartic space curve, shown with the Voronoi cell in one of its normal planes.

The distance from the variety to its medial axis, which is the answer to the “how far” question, is called the *reach* of  $X$ . This quantity is of interest, for example, in topological data analysis, as it is the main quantity determining the density of sample points needed to compute the persistent homology of  $X$ . We refer to [31, 68] for studies on sampling at the interface of topological data analysis with metric algebraic geometry. The distance from a point  $\mathbf{y}$  on  $X$  to the variety’s medial axis could be considered the *local reach* of  $X$ . Equivalently, this is the distance from  $\mathbf{y}$  to the boundary of its Voronoi cell  $\text{Vor}_X(\mathbf{y})$ .

The material that follows is based on the article [52]. We begin with the exact symbolic computation of the Voronoi boundary at  $\mathbf{y}$  from the equations that define  $X$ . This uses a Gröbner-based algorithm whose input is  $\mathbf{y}$  and the ideal of  $X$  and whose output is the ideal defining  $\partial_{\text{alg}}\text{Vor}_X(\mathbf{y})$ . In the next section we present formulas for the Voronoi degree  $\delta_X(\mathbf{y})$  when  $X$  and  $\mathbf{y}$  are sufficiently general and  $\dim(X) \leq 2$ . The proofs of these formulas require some intersection theory. Thereafter we study the case when  $\mathbf{y}$  is a low rank matrix and  $X$  is the variety of these matrices. This relies on the *Eckart-Young Theorem*.

We now describe Gröbner basis methods for finding the Voronoi boundaries of a given variety. We start with an ideal  $I = \langle f_1, f_2, \dots, f_m \rangle$  in  $\mathbb{Q}[x_1, \dots, x_n]$  whose real variety  $X = V(I) \subset \mathbb{R}^n$  is assumed to be nonempty. We assume that  $I$  is real radical and prime, so that  $X_{\mathbb{C}}$  is an irreducible variety in  $\mathbb{C}^n$  whose real points are Zariski dense. Our aim is to compute the Voronoi boundary of a given point  $\mathbf{y} \in X$ . In our examples, the coordinates of the point  $\mathbf{y}$  and the coefficients of the polynomials  $f_i$  are rational numbers. Under these assumptions, the following computations can be done in polynomial rings over  $\mathbb{Q}$ .

Fix the polynomial ring  $R = \mathbb{Q}[x_1, \dots, x_n, u_1, \dots, u_n]$  where  $\mathbf{u} = (u_1, \dots, u_n)$  is an auxiliary point with unknown coordinates. The *augmented Jacobian* of  $X$  at  $\mathbf{x}$  is the following matrix of size  $(m+1) \times n$  with entries in  $R$ . It contains the  $n$  partial derivatives of the  $m$  generators of  $I$ :

$$J_I(\mathbf{x}, \mathbf{u}) := \begin{bmatrix} \mathbf{u} - \mathbf{x} \\ (\nabla f_1)(\mathbf{x}) \\ \vdots \\ (\nabla f_m)(\mathbf{x}) \end{bmatrix}$$

Let  $N_I$  denote the ideal in  $R$  generated by  $I$  and the  $(c+1) \times (c+1)$  minors of the augmented Jacobian  $J_I(\mathbf{x}, \mathbf{u})$ , where  $c$  is the codimension of the given variety  $X \subset \mathbb{R}^n$ . The ideal  $N_I$  in  $R$  defines a subvariety

of dimension  $n$  in  $\mathbb{R}^{2n}$ , namely the *Euclidean normal bundle* of  $X$ . Its points are pairs  $(\mathbf{x}, \mathbf{u})$  where  $\mathbf{x}$  is a point in the given variety  $X$  and  $\mathbf{u}$  lies in the normal space of  $X$  at  $\mathbf{x}$ .

**Example 8.5 (Cuspidal cubic)** Let  $n = 2$  and  $I = \langle x_1^3 - x_2^2 \rangle$ , so  $X = V(I) \subset \mathbb{R}^2$  is a cubic curve with a cusp at the origin. The ideal of the Euclidean normal bundle of  $X$  is generated by two polynomials:

$$N_I = \langle x_1^3 - x_2^2, \det \begin{pmatrix} u_1 - x_1 & u_2 - x_2 \\ 3x_1^2 & -2x_2 \end{pmatrix} \rangle \subset R = \mathbb{Q}[x_1, x_2, u_1, u_2]$$

For any  $\mathbf{y} \in X$ , let  $N_I(\mathbf{y})$  denote the linear ideal that is obtained from  $N_I$  by replacing the unknown point  $\mathbf{x}$  by the specific point  $\mathbf{y}$ . For instance, if that point is  $\mathbf{y} = (4, 8)$  then  $N_I(\mathbf{y}) = \langle u_1 + 3u_2 - 28 \rangle$ .

Returning to the general setting, we define the *critical ideal* of the variety  $X$  at the point  $\mathbf{y}$  as

$$C_I(\mathbf{y}) = N_I + N_I(\mathbf{y}) + \langle \|\mathbf{x} - \mathbf{u}\|^2 - \|\mathbf{y} - \mathbf{u}\|^2 \rangle \subset R.$$

The variety of the ideal  $C_I(\mathbf{y})$  consists of pairs  $(\mathbf{u}, \mathbf{x})$  such that  $\mathbf{x}$  and  $\mathbf{y}$  are equidistant from  $\mathbf{u}$  and both are critical points of the distance function from  $\mathbf{u}$  to  $X$ . The *Voronoi ideal* is the following ideal in  $\mathbb{Q}[u_1, \dots, u_n]$ . It is obtained from the critical ideal by saturation and elimination:

$$\text{Vor}_I(\mathbf{y}) = (C_I(\mathbf{y}) : \langle \mathbf{x} - \mathbf{y} \rangle^\infty) \cap \mathbb{Q}[u_1, \dots, u_n]. \quad (8.3)$$

The geometric interpretation of each step in our construction implies the following result:

**Proposition 8.6** *The affine variety in  $\mathbb{C}^n$  defined by the Voronoi ideal  $\text{Vor}_I(\mathbf{y})$  contains the algebraic Voronoi boundary  $\partial_{\text{alg}} \text{Vor}_X(\mathbf{y})$  of the given real variety  $X$  at its point  $\mathbf{y}$ .*

*Remark 8.7* The verb “contains” sounds weak, but it is much stronger than it may seem. Indeed, in generic situations, the ideal  $\text{Vor}_I(\mathbf{y})$  will be prime, and it defines an irreducible hypersurface in the normal space  $N_I(\mathbf{y})$ . This hypersurface equals the algebraic Voronoi boundary, so containment is an equality. We saw this in Example 8.3. For special data,  $\text{Vor}_I(\mathbf{y})$  usually defines a hypersurface in  $N_I(\mathbf{y})$ , but it can have extraneous components, which are often easy to remove.

**Example 8.8** For the point  $\mathbf{y} = (4, 8)$  on the cuspidal cubic  $X$  in Example 8.5, we have  $N_I(\mathbf{y}) = \langle u_1 + 3u_2 - 28 \rangle$ . Going through the steps above, we find that the Voronoi ideal is

$$\text{Vor}_I(\mathbf{y}) = \langle u_1 - 28, u_2 \rangle \cap \langle u_1 + 26, u_2 - 18 \rangle \cap \langle u_1 + 3u_2 - 28, 27u_2^2 - 486u_2 + 2197 \rangle.$$

The third component has no real roots and is hence extraneous. The Voronoi boundary consists of two points. Namely, we have  $\partial \text{Vor}_X(\mathbf{y}) = \{(28, 0), (-26, 18)\}$ . The Voronoi cell  $\text{Vor}_X(\mathbf{y})$  is the line segment connecting these points. This segment is shown in green in Figure 8.2. Its right endpoint  $(28, 0)$  is equidistant from  $\mathbf{y}$  and the point  $(4, -8)$ . Its left endpoint  $(-26, 18)$  is equidistant from  $\mathbf{y}$  and the origin  $(0, 0)$ , whose Voronoi cell will be discussed in Remark 8.9.

The cuspidal cubic  $X$  is very special. If we replace  $X$  by a general cubic (defined over  $\mathbb{Q}$ ) in the affine plane, then  $\text{Vor}_I(\mathbf{y})$  is generated modulo  $N_I(\mathbf{y})$  by an irreducible polynomial of degree eight in  $\mathbb{Q}[u_2]$ . Thus, the expected Voronoi degree for general (affine) plane cubics is  $\delta_X(\mathbf{y}) = 8$ .

*Remark 8.9 (Singularities)* Voronoi cells at singular points can be computed with the same procedure as above. However, these Voronoi cells generally have higher dimensions. For an illustration, consider the cuspidal cubic, and let  $\mathbf{y} = (0, 0)$  be the cusp. A Gröbner basis computation yields the Voronoi boundary  $27u_2^4 + 128u_1^3 + 72u_1u_2^2 + 32u_1^2 + u_2^2 + 2u_1$ . The Voronoi cell is the two-dimensional convex region bounded by this quartic, shown in blue in Figure 8.2. The Voronoi cell might also be empty at a singularity. This happens for instance for  $V(x_1^3 + x_1^2 - x_2^2)$ , which has an ordinary double point at  $\mathbf{y} = (0, 0)$ . In general, the cell dimension depends on both the embedding dimension and the branches of the singularity.

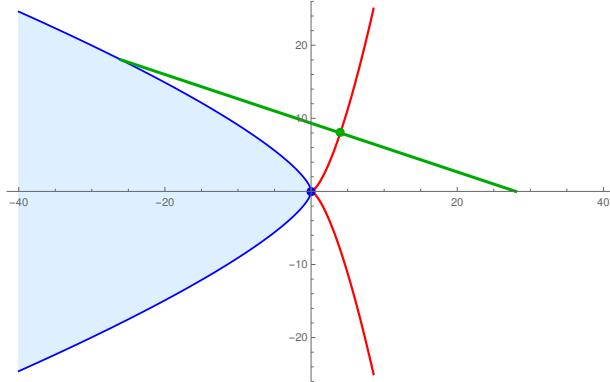


Fig. 8.2: The cuspidal cubic is shown in red. The Voronoi cell of a smooth point is a green line segment. The Voronoi cell of the cusp is the convex region bounded by the blue curve.

Proposition 8.6 gives an algorithm for computing the Voronoi ideal \$\text{Vor}\_I(\mathbf{y})\$ when \$\mathbf{y}\$ is a smooth point in \$X = V(I)\$. Experiments with **Macaulay2** [79] are reported in [52]. For small enough instances, the computation terminates and we obtain the defining polynomial of the Voronoi boundary \$\partial\_{\text{alg}}\text{Vor}\_X(\mathbf{y})\$. This polynomial is unique modulo the linear ideal of the normal space \$N\_I(\mathbf{y})\$. For larger instances, we can only compute the degree of \$\partial\_{\text{alg}}\text{Vor}\_X(\mathbf{y})\$ but not its equation. This is done by working over a finite field and adding \$c - 1\$ random linear equations in \$u\_1, \dots, u\_n\$ in order to get a zero-dimensional polynomial system.

Computations are easiest to set up for the case of hypersurfaces (\$c = 1\$). One can explore random polynomials \$f\$ of degree \$d\$ in \$\mathbb{Q}[x\_1, \dots, x\_n]\$, both inhomogeneous and homogeneous. These are chosen among those that vanish at a preselected point \$\mathbf{y}\$ in \$\mathbb{Q}^n\$. In each iteration, the Voronoi ideal \$\text{Vor}\_I(\mathbf{y})\$ from (8.3) was found to be zero-dimensional. In fact, \$\text{Vor}\_I(\mathbf{y})\$ is a maximal ideal in \$\mathbb{Q}[u\_1, \dots, u\_n]\$, and the Voronoi degree \$\delta\_X(\mathbf{y})\$ is the degree of the field extension of \$\mathbb{Q}\$ that is defined by that maximal ideal.

We summarize our results in Tables 8.1 and 8.2, and we extract conjectural formulas.

| $n \setminus d$ | 2 | 3   | 4    | 5   | 6    | 7   | 8  | $\delta_X(\mathbf{y}) = \text{degree}(\text{Vor}_{(f)}(\mathbf{y}))$ |
|-----------------|---|-----|------|-----|------|-----|----|--|
| 1               | 1 | 2   | 3    | 4   | 5    | 6   | 7  | $d-1$  |
| 2               | 2 | 8   | 16   | 26  | 38   | 52  | 68 | $d^2+d-4$  |
| 3               | 3 | 23  | 61   | 123 | 215  | 343 |    | $d^3+d-7$  |
| 4               | 4 | 56  | 202  | 520 | 1112 |     |    | $d^4-d^3+d^2+d-10$   |
| 5               | 5 | 125 | 631  |     |      |     |    | $d^5-2d^4+2d^3+d-13$   |
| 6               | 6 | 266 | 1924 |     |      |     |    | $d^6-3d^5+4d^4-2d^3+d^2+d-16$  |
| 7               | 7 | 551 |      |     |      |     |    | $d^7-4d^6+7d^5-6d^4+3d^3+d-19$                                       |

Table 8.1: The Voronoi degree of an inhomogeneous polynomial \$f\$ of degree \$d\$ in \$\mathbb{R}^n\$.

*Conjecture 8.10* The Voronoi degree of a generic hypersurface of degree \$d\$ in \$\mathbb{R}^n\$ equals

$$(d-1)^n + 3(d-1)^{n-1} + \frac{4}{d-2}((d-1)^{n-1} - 1) - 3n.$$

The Voronoi degree of the cone of a generic homogeneous polynomial of degree \$d\$ in \$\mathbb{R}^n\$ is

$$2(d-1)^{n-1} + \frac{4}{d-2}((d-1)^{n-1} - 1) - 3n + 2.$$

| $n \setminus d$ | 2 | 3   | 4   | 5   | 6  | 7  | 8   | $\delta_X(y) = \text{degree}(\text{Vor}_{\langle f \rangle}(y))$ |
|-----------------|---|-----|-----|-----|----|----|-----|--|
| 2               | 2 | 4   | 6   | 8   | 10 | 12 | 14  | $2d-2$   |
| 3               | 3 | 13  | 27  | 45  | 67 | 93 | 123 | $2d^2-5$   |
| 4               | 4 | 34  | 96  | 202 |    |    |     | $2d^3-2d^2+2d-8$   |
| 5               | 5 | 79  | 309 |     |    |    |     | $2d^4-4d^3+4d^2-11$  |
| 6               | 6 | 172 |     |     |    |    |     | $2d^5-6d^4+8d^3-4d^2+2d-14$                                      |
| 7               | 7 | 361 |     |     |    |    |     | $2d^6-8d^5+14d^4-12d^3+6d^2-17$                                  |

Table 8.2: The Voronoi degree of a homogeneous polynomial  $f$  of degree  $d$  in  $\mathbb{R}^n$ .

Both parts of this conjecture are proved for  $n \leq 3$  in [52, Section 4], where the geometric theory of Voronoi degrees of low-dimensional varieties is developed. The case  $d = 2$  was analyzed in [51, Proposition 5.8]. In general, for  $n \geq 4$  and  $d \geq 3$ , the problem is open.

### 8.3 Formulas from Algebraic Geometry

To recap, the algebraic boundary of the Voronoi cell  $\text{Vor}_X(\mathbf{y})$  is a hypersurface in the normal space to a variety  $X \subset \mathbb{R}^n$  at a point  $\mathbf{y} \in X$ . We shall present formulas for the degree  $\delta_X(\mathbf{y})$  of that hypersurface when  $X$  is a curve or a surface. All proofs appear in [52, Section 6]. We identify  $X$  and  $\partial_{\text{alg}}\text{Vor}_X(\mathbf{y})$  with their Zariski closures in complex projective space  $\mathbb{P}^n$ , so there is a natural assigned hyperplane at infinity. We say that  $X$  is in *general position* in  $\mathbb{P}^n$  if the hyperplane at infinity intersects  $X$  transversally, i.e. that the intersection is smooth.

**Theorem 8.11** *Let  $X \subset \mathbb{P}^n$  be a curve of degree  $d$  and geometric genus  $g$  with at most ordinary multiple points as singularities. The Voronoi degree at a general point  $\mathbf{y} \in X$  equals*

$$\delta_X(\mathbf{y}) = 4d + 2g - 6,$$

*provided  $X$  is in general position in  $\mathbb{P}^n$ .*

**Example 8.12** If  $X$  is a smooth curve of degree  $d$  in the plane, then  $2g - 2 = d(d - 3)$ , so

$$\delta_X(\mathbf{y}) = d^2 + d - 4.$$

This confirms our experimental results in the row  $n = 2$  of Table 8.1.

**Example 8.13** If  $X$  is a rational curve of degree  $d$ , then  $g = 0$  and hence  $\delta_X(\mathbf{y}) = 4d - 6$ . If  $X$  is an elliptic curve, so the genus is  $g = 1$ , then we have  $\delta_X(\mathbf{y}) = 4d - 4$ . A space curve with  $d = 4$  and  $g = 1$  was studied in Example 8.4. Its Voronoi degree equals  $\delta_X(\mathbf{y}) = 12$ .

Theorem 8.11 is [52, Theorem 5.1]. The general position assumption is essential. For an example, let  $X$  be the twisted cubic curve in  $\mathbb{P}^3$ , with affine parameterization  $t \mapsto (t, t^2, t^3)$ . Here  $g = 0$  and  $d = 3$ , so the expected Voronoi degree is 6. However, a computation shows that  $\delta_X(\mathbf{y}) = 4$ . This drop arises because the plane at infinity in  $\mathbb{P}^3$  intersects the curve  $X$  in a triple point. After a general linear change of coordinates in  $\mathbb{P}^3$ , which amounts to a linear fractional transformation in  $\mathbb{R}^3$ , we correctly find  $\delta_X(\mathbf{y}) = 6$ .

We next present a formula for the Voronoi degree of a surface  $X$  which is smooth and irreducible in  $\mathbb{P}^n$ . Our formula is in terms of its degree  $d$  and two further invariants. The first, denoted  $\chi(X)$ , is the topological Euler characteristic. This equals the degree of the second Chern class of the tangent bundle. The second invariant, denoted  $g(X)$ , is the genus of the curve obtained by intersecting  $X$  with a general quadratic hypersurface in  $\mathbb{P}^n$ . Thus,  $g(X)$  is the quadratic analogue to the sectional genus of the surface  $X$ .

**Theorem 8.14 (Theorem 5.4 in [52])** Let  $X \subset \mathbb{P}^n$  be a smooth surface of degree  $d$ . Then

$$\delta_X(y) = 3d + \chi(X) + 4g(X) - 11,$$

provided the surface  $X$  is in general position in  $\mathbb{P}^n$  and  $y$  is a general point on  $X$ .

**Example 8.15** If  $X$  is a smooth surface in  $\mathbb{P}^3$  of degree  $d$ , then  $\chi(X) = d(d^2 - 4d + 6)$ , by [73, Ex 3.2.12]. A smooth quadratic hypersurface section of  $X$  is an irreducible curve of degree  $(d, d)$  in  $\mathbb{P}^1 \times \mathbb{P}^1$ . The genus of such a curve is  $g(X) = (d - 1)^2$ . We conclude that

$$\delta_X(y) = 3d + d(d^2 - 4d + 6) + 4(d - 1)^2 - 11 = d^3 + d - 7.$$

This confirms our experimental results in the row  $n = 3$  of Table 8.1.

**Example 8.16** Let  $X$  be the Veronese surface of order  $e$  in  $\mathbb{P}^{(e+1)/2-1}$ , taken after a general linear change of coordinates in that ambient space. The degree of  $X$  equals  $d = e^2$ . We have  $\chi(X) = \chi(\mathbb{P}^2) = 3$ , and the general quadratic hypersurface section of  $X$  is a curve of genus  $g(X) = \binom{2e-1}{2}$ . We conclude that the Voronoi degree of  $X$  at a general point  $y$  equals

$$\delta_X(y) = 3e^2 + 3 + 2(2e-1)(2e-2) - 11 = 11e^2 - 12e - 4.$$

For instance, for the quadratic Veronese surface in  $\mathbb{P}^5$  we have  $e = 2$  and hence  $\delta_X(y) = 16$ . This is smaller than the number 18 found in Example 8.22, since back then we were dealing with the cone over the Veronese surface in  $\mathbb{R}^6$ , and not with the Veronese surface in  $\mathbb{R}^5 \subset \mathbb{P}^5$ .

We finally consider affine surfaces defined by homogeneous polynomials. Namely, let  $X \subset \mathbb{R}^n$  be the affine cone over a general smooth curve of degree  $d$  and genus  $g$  in  $\mathbb{P}^{n-1}$ .

**Theorem 8.17 (Theorem 5.7 in [52])** If  $X \subset \mathbb{R}^n$  is the cone over a smooth curve in  $\mathbb{P}^{n-1}$  then

$$\delta_X(y) = 6d + 4g - 9,$$

provided that the curve is in general position and  $y$  is a general point.

**Example 8.18** If  $X \subset \mathbb{R}^3$  is the cone over a smooth curve of degree  $d$  in  $\mathbb{P}^2$ , then  $2g - 2 = d(d - 3)$ , by the degree-genus formula for plane curves. We conclude that the Voronoi degree of  $X$  is equal to

$$\delta_X(y) = 2d^2 - 5.$$

This confirms our experimental results in the row  $n = 3$  of Table 8.2.

Let us comment on the assumptions made in our theorems. We assumed that the variety  $X$  is in general position in  $\mathbb{P}^n$ . If this is not satisfied, then the Voronoi degree may drop. The point here is that the Voronoi ideal  $\text{Vor}_I(y)$  depends polynomially on the description of  $X$ , and the degree of this zero-dimensional ideal can only go down – and not up – when that description specializes. Making this statement precise would require a technical discussion of families in algebraic geometry, a topic best left to the experts on foundations. Nonetheless, the technique introduced in the next section can be adapted to determine the correct value. As an illustration, we consider the affine Veronese surface (Example 8.16).

**Example 8.19** Let  $X \subset \mathbb{P}^5$  be the Veronese surface with affine parametrization  $(s, t) \mapsto (s, t, s^2, st, t^2)$ . The hyperplane at infinity intersects  $X$  in a double conic, so  $X$  is not in general position. In the next section, we will show that the true Voronoi degree is  $\delta_X(y) = 10$ . For the Frobenius norm, the Voronoi degree drops further. For this, we shall derive  $\delta_X(y) = 4$ .

## 8.4 Voronoi meets Eckhart-Young

We now turn to the case of great interest in applications. Let  $X$  be the variety of real  $m \times n$  matrices of rank  $\leq r$ . We consider two natural norms on the space  $\mathbb{R}^{m \times n}$  of real  $m \times n$  matrices. Our first matrix norm is the *Frobenius norm*  $\|U\|_F := \sqrt{\sum_{ij} U_{ij}^2}$ . Our second matrix norm is the *spectral norm*  $\|U\|_2 := \max_i \sigma_i(U)$  which extracts the largest singular value of the matrix  $U$ .

Fix a rank  $r$  matrix  $V$  in  $X$ . This is a nonsingular point in  $X$ . We consider the Voronoi cell  $\text{Vor}_X(V)$  with respect to the Frobenius norm. This is consistent with our setting because the Frobenius norm agrees with the Euclidean norm on  $\mathbb{R}^{m \times n}$ . This identification will no longer be valid when we restrict to the subspace of symmetric matrices.

Fix  $U \in \text{Vor}_X(V)$ . This means that the closest point to  $U$  in the rank  $r$  variety  $X$  is the matrix  $V$ . By the Eckart-Young Theorem, the matrix  $V$  is derived from  $U$  by computing the singular value decomposition  $U = \Sigma_1 D \Sigma_2$ . Here  $\Sigma_1$  and  $\Sigma_2$  are orthogonal matrices of size  $m \times m$  and  $n \times n$  respectively, and  $D$  is a nonnegative diagonal matrix whose entries are the singular values. Let  $D^{[r]}$  be the matrix that is obtained from  $D$  by replacing all singular values except for the  $r$  largest ones by zero. Then, according to Eckart-Young, we have  $V = \Sigma_1 \cdot D^{[r]} \cdot \Sigma_2$ .

*Remark 8.20* The Eckart-Young Theorem works for both the Frobenius norm and the spectral norm. This means that  $\text{Vor}_X(V)$  is also the Voronoi cell for the spectral norm.

The following theorem describes the Voronoi cells for low-rank matrix approximation.

**Theorem 8.21** *Let  $V$  be an  $m \times n$ -matrix of rank  $r$ . The Voronoi cell  $\text{Vor}_X(V)$  is congruent up to scaling to the unit ball in the spectral norm on the space of  $(m-r) \times (n-r)$ -matrices.*

Before we present the proof, let us first see why the statement makes sense. The determinantal variety  $X$  has dimension  $rm + rn - r^2$  in an ambient space of dimension  $mn$ . The dimension of the normal space at a point is the difference of these two numbers, so it equals  $(m-r)(n-r)$ . Every Voronoi cell is a full-dimensional convex body in the normal space. Next consider the case  $m = n$  and restrict to the space of diagonal matrices. Now  $X$  is the set of vectors in  $\mathbb{R}^n$  having at most  $r$  nonzero coordinates. This is a reducible variety with  $\binom{n}{r}$  components, each a coordinate subspace. For a general point  $y$  in such a subspace, the Voronoi cell  $\text{Vor}_X(y)$  is a convex polytope. It is congruent to a regular cube of dimension  $n-r$ , which is the unit ball in the  $L^\infty$ -norm on  $\mathbb{R}^{n-r}$ . Theorem 8.21 describes the orbit of this picture under the action of the two orthogonal groups on  $\mathbb{R}^{m \times n}$ .

For example, consider the special case where  $n = 3$  and  $r = 1$ . In this case,  $X$  consists of the three coordinate axes in  $\mathbb{R}^3$ . The Voronoi decomposition of this reducible curve decomposes  $\mathbb{R}^3$  into squares, each normal to a different point on the three lines. The image of this picture under orthogonal transformations is the Voronoi decomposition of  $\mathbb{R}^{3 \times 3}$  associated with the affine variety of rank 1 matrices. That variety has dimension 5, and each Voronoi cell is a 4-dimensional convex body in the normal space.

**Proof (of Theorem 8.21)** The Voronoi cell is invariant under orthogonal transformations. We may therefore assume that the matrix  $V = (v_{ij})$  satisfies  $v_{11} \geq v_{22} \geq \dots \geq v_{rr} = u > 0$  and  $v_{ij} = 0$  for all other entries. The Voronoi cell of the diagonal matrix  $V$  consists of matrices  $U$  whose block-decomposition into  $r + (m-r)$  rows and  $r + (n-r)$  columns satisfies

$$\begin{pmatrix} I & 0 \\ 0 & T_1 \end{pmatrix} \cdot \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix} \cdot \begin{pmatrix} I & 0 \\ 0 & T_2 \end{pmatrix} = \begin{pmatrix} V_{11} & 0 \\ 0 & V_{22} \end{pmatrix}.$$

Here  $V_{11} = \text{diag}(v_{11}, \dots, v_{rr})$  agrees with the upper  $r \times r$ -block of  $V$ , and  $V_{22}$  is a diagonal matrix whose entries are bounded above by  $u$  in absolute value. This implies that  $U_{11} = V_{11}$ ,  $U_{12} = 0$  and  $U_{21} = 0$ ,

Moreover,  $U_{22}$  is an arbitrary  $(m-r) \times (n-r)$  matrix with spectral norm at most  $u$ . Hence the Voronoi cell of the given diagonal matrix  $V$  is congruent to the set of all such matrices  $U_{22}$ . This convex body equals  $u$  times the unit ball in  $\mathbb{R}^{(m-r) \times (n-r)}$  under the spectral norm.  $\square$

Our problem becomes even more interesting when we restrict to matrices in a linear subspace. To see this, let  $X$  denote the variety of symmetric  $n \times n$  matrices of rank  $\leq r$ . We can regard  $X$  either as a variety in the ambient matrix space  $\mathbb{R}^{n \times n}$ , or in the space  $\mathbb{R}^{\binom{n+1}{2}}$  whose coordinates are the upper triangular entries of a symmetric matrix. On the latter space we have both the *Euclidean norm* and the *Frobenius norm*. These are now different!

The Frobenius norm on  $\mathbb{R}^{\binom{n+1}{2}}$  is the restriction of the Frobenius norm on  $\mathbb{R}^{n \times n}$  to the subspace of symmetric matrices. For instance, if  $n = 2$ , we identify the vector  $(a, b, c)$  with the symmetric matrix  $\begin{pmatrix} a & b \\ b & c \end{pmatrix}$ . The Frobenius norm of this matrix is  $\sqrt{a^2 + 2b^2 + c^2}$ , whereas the Euclidean norm is  $\sqrt{a^2 + b^2 + c^2}$ . The two norms have dramatically different properties with respect to low rank approximation. The Eckart-Young Theorem remains valid for the Frobenius norm on  $\mathbb{R}^{\binom{n+1}{2}}$ , but it is not valid for the Euclidean norm. The implications of this are explained in [62, Example 3.2].

In what follows we elucidate this point by comparing the Voronoi cells with respect to the two norms.

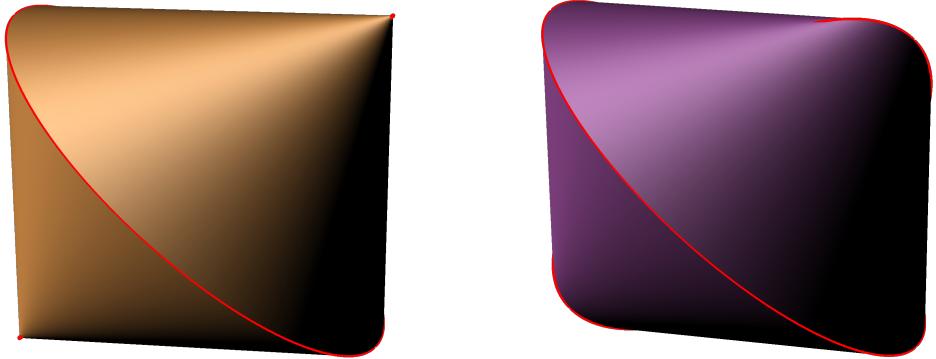


Fig. 8.3: The Voronoi cell of a symmetric  $3 \times 3$  matrix of rank 1 is a convex body of dimension 3. It is shown for the Frobenius norm (left) and for the Euclidean norm (right).

**Example 8.22** Let  $X$  be the variety of symmetric  $3 \times 3$  matrices of rank  $\leq 1$ . For the Euclidean metric,  $X$  lives in  $\mathbb{R}^6$ . For the Frobenius metric,  $X$  lives in a 6-dimensional subspace of  $\mathbb{R}^{3 \times 3}$ . Let  $V$  be a smooth point in  $X$ , i.e. a symmetric  $3 \times 3$  matrix of rank 1. The normal space to  $X$  at  $V$  has dimension 3. Hence, in either norm, the Voronoi cell  $\text{Vor}_X(V)$  is a 3-dimensional convex body. Figure 8.3 illustrates these two bodies.

For the Frobenius metric, the Voronoi cell is congruent to the set of matrices  $\begin{pmatrix} a & b \\ b & c \end{pmatrix}$  with eigenvalues between  $-1$  and  $1$ . This semialgebraic set is bounded by the surfaces defined by the singular quadrics  $\det \begin{pmatrix} a+1 & b \\ b & c+1 \end{pmatrix}$  and  $\det \begin{pmatrix} a-1 & b \\ b & c-1 \end{pmatrix}$ . The Voronoi ideal is of degree 4, defined by the product of these two determinants (modulo the normal space). The Voronoi cell is shown on the left in Figure 8.3. It is the intersection of two quadratic cones. The cell is the convex hull of the circle in which the two quadrics meet, together with the two vertices.

For the Euclidean metric, the Voronoi boundary at a generic point  $V$  in  $X$  is defined by an irreducible polynomial of degree 18 in  $a, b, c$ . In some cases, the Voronoi degree can drop. For instance, consider the special rank 1 matrix  $V = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ . For this point, the degree of the Voronoi boundary is only 12. This

particular Voronoi cell is shown on the right in Figure 8.3. This cell is the convex hull of two ellipses, which are shown in red in the diagram.



## **Chapter 9**

### **Condition Numbers**

The concept of a condition number has its origin in numerical analysis. It measures how much the output value of a function we wish to evaluate can change for a small change in the input argument. In this chapter we discuss condition numbers in the context of metric algebraic geometry. In the first section we offer an introduction to the relevant notions for assessing errors in numerical computations.

Suppose the function to be evaluated is an algebraic function. For instance, we may wish to map the coefficients of a univariate polynomial to one of its roots. This function is well-defined locally, and it is well-behaved when the polynomial is far from the hypersurface defined by the discriminant. Indeed, the condition number stands in a reciprocal relationship to the distance to the variety of ill-posed instances.

This variety of ill-posed problems is often a discriminantal hypersurface, and thus we are naturally led to the ED problem for discriminants. This topic will be studied in the third and last section of this chapter. The classical discriminant is the dual variety to the Veronese variety, and other discriminants are dual to other toric varieties. We can apply ED duality (Theorem 2.23) to gain insights and computational speed.

A special case is the determinant of a square matrix, which arises when our function is matrix inversion. The relevant ED problem points us to the Eckhard-Young Theorem (Theorem 2.10) and this is why we include the proof of Eckhard-Young in the second section, about the condition number of matrix inversion.

## 9.1 Errors in Numerical Computations

Input data for numerical algorithms can have errors, caused, for instance, by measurements errors. Hence, the output of the computation also has errors. We wish to compare the output error with the input error.

**Example 9.1 (Exact algorithm)** Given a matrix  $A \in \mathbb{R}^{2 \times 2}$  with  $\det(A) \neq 0$ , we want to compute its inverse. We consider two instances of this problem. The errors are measured by the Euclidean norm.

1. First, let  $A = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$ . We consider this matrix to be the true input data. A small measurement error gives the new input data  $\tilde{A} = \begin{pmatrix} 1 & 1 \\ -1+\varepsilon & 1 \end{pmatrix}$ , where  $0 < \varepsilon \ll 1$ . The *exact* solutions  $A^{-1}$  and  $\tilde{A}^{-1}$  are then

$$A^{-1} = \frac{1}{2} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}, \quad \text{and} \quad \tilde{A}^{-1} = \frac{1}{2-\varepsilon} \begin{pmatrix} 1 & -1 \\ 1-\varepsilon & 1 \end{pmatrix} = A^{-1} + \frac{\varepsilon}{2(2-\varepsilon)} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

Comparing the errors, we find that  $\|A^{-1} - \tilde{A}^{-1}\| \approx \|A - \tilde{A}\|$ . In words, the error in the input  $\|A - \tilde{A}\|$  and the error in the output  $\|A^{-1} - \tilde{A}^{-1}\|$  are roughly the same. They both are of the order  $O(\varepsilon)$ .

2. The true input in the second example is the matrix  $B = \begin{pmatrix} 1 & 1 \\ 1 & 1+\delta \end{pmatrix}$ , where  $|\delta| \neq 0$  is small. We perturb the input by adding  $\varepsilon$  to the lower left entry. The perturbed input is  $\tilde{B} = \begin{pmatrix} 1 & 1 \\ 1+\varepsilon & 1+\delta \end{pmatrix}$ . The matrix inverses are

$$B^{-1} = \frac{1}{\delta} \begin{pmatrix} 1+\delta & -1 \\ -1 & 1 \end{pmatrix} \quad \text{and} \quad \tilde{B}^{-1} = \frac{1}{\delta-\varepsilon} \begin{pmatrix} 1+\delta & -1 \\ -1-\varepsilon & 1 \end{pmatrix} = B^{-1} + \frac{\varepsilon}{\delta(\varepsilon-\delta)} \begin{pmatrix} 1+\delta & -1 \\ -(1+\delta) & 1 \end{pmatrix}.$$

This implies  $\|B^{-1} - \tilde{B}^{-1}\| \approx \frac{1}{\delta(\varepsilon-\delta)} \cdot \|B - \tilde{B}\|$ . If  $\varepsilon < \delta$ , then we have an amplification of the error by a factor of roughly  $\delta^{-2}$ , which is large. The behavior here is different from that before.

We applied an exact algorithm to the problem, but we observed considerable differences in the output. In the first example, the output  $\tilde{A}^{-1}$  for the perturbed data  $\tilde{A}$  was close the true output  $A^{-1}$ . On the other hand, in the second example the output  $\tilde{B}^{-1}$  for the perturbed data  $\tilde{B}$  was far from the true output  $B^{-1}$ .  $\diamond$

The previous example shows that, even if we can compute the *exact* solution of a problem, small errors in the data may be amplified tremendously in the output. The theory of *condition numbers* helps us to understand when and why this happens. In simple terms, a condition number is a quantity associated to a *computational problem*, and it measures the sensitivity of the output to small errors in the input data.

**Definition 9.2** A *computational problem* is a function  $\phi : M \rightarrow N$  from a space  $M$  of inputs to a space  $N$  of outputs. For us, each space is a subset of a Euclidean space, and it carries the induced Euclidean metric.

**Example 9.3** In Example 9.1 the input space and the output space is  $M = N = \{A \in \mathbb{R}^{2 \times 2} \mid \det(A) \neq 0\}$ . The computational problem is matrix inversion, so the relevant function is  $\phi(A) = A^{-1}$ .  $\diamond$

We regard  $(M, d_N)$  and  $(N, d_N)$  as metric spaces. The following definition is due to Rice [156].

**Definition 9.4** The (*absolute*) *condition number* of  $\phi : M \rightarrow N$  at the input datum  $\mathbf{x} \in M$  is

$$\kappa[\phi](\mathbf{x}) := \lim_{\varepsilon \rightarrow 0} \sup_{\substack{\mathbf{y} \in M: \\ d_M(\mathbf{x}, \mathbf{y}) \leq \varepsilon}} \frac{d_N(\phi(\mathbf{x}), \phi(\mathbf{y}))}{d_M(\mathbf{x}, \mathbf{y})}.$$

The motivation for this definition of a condition number is as follows: For small  $d_M(\mathbf{x}, \mathbf{y})$  we have

$$d_N(\phi(\mathbf{x}), \phi(\mathbf{y})) \leq \kappa[\phi](\mathbf{x}) \cdot d_M(\mathbf{x}, \mathbf{y}) + o(d_M(\mathbf{x}, \mathbf{y})).$$

In words, a small error  $\varepsilon = d_M(\mathbf{x}, \mathbf{y})$  in the input data causes an error of roughly  $\kappa[\phi](\mathbf{x}) \cdot \varepsilon$  in the output data. This is entirely independent of the algorithm which is used to evaluate  $\phi(\mathbf{x})$ . At this stage, given the lim-sup definition, it is unclear how condition numbers can be computed. As we shall see, this is where the geometric perspective of Demmel [57] comes in. But, let's first discuss an important variant.

*Remark 9.5* Fix the Euclidean spaces  $M = \mathbb{R}^n$  and  $N = \mathbb{R}^m$ . What does “small error” mean in this case? If  $\|\mathbf{x}\| = 10^4$ , is an error of size  $\|\mathbf{x} - \mathbf{y}\| = 10^2$  small or large? To address a question like this, there is a notion of relative error in numerical analysis. By definition, the *relative error* is

$$\text{RelError}(\mathbf{x}, \mathbf{y}) = \frac{\|\mathbf{x} - \mathbf{y}\|}{\|\mathbf{x}\|} \quad \text{for } \mathbf{x}, \mathbf{y} \in M.$$

With this, we define the *relative condition number* of the function  $\phi$  at the input datum  $\mathbf{x} \in M$  as follows:

$$\kappa_{\text{REL}}[\phi](\mathbf{x}) := \lim_{\varepsilon \rightarrow 0} \sup_{\text{RelError}(\mathbf{x}, \mathbf{y}) \leq \varepsilon} \frac{\text{RelError}(\phi(\mathbf{x}), \phi(\mathbf{y}))}{\text{RelError}(\mathbf{x}, \mathbf{y})} = \kappa[\phi](\mathbf{x}) \cdot \frac{\|\mathbf{x}\|_M}{\|\phi(\mathbf{x})\|_N}.$$

In numerical analysis, relative errors are more significant than absolute errors, because *floating point* arithmetic introduces relative errors (see, e.g., [93] or [178, p. 91]). Modern architecture is optimized for computing with floating point numbers. A floating point number system  $\mathcal{F}$  is a subset of the real numbers  $\mathbb{R}$  that is specified by four integers  $\beta, t, e_{\min}, e_{\max}$ , where  $\beta$  is called *base*,  $t$  is called *precision*, and  $[e_{\min}, e_{\max}]$  is called *exponential range*. Then,  $\mathcal{F} = \{\pm \beta^e \sum_{i=1}^t \frac{d_i}{\beta} \mid 0 \leq d_i \leq \beta-1, e_{\min} \leq e \leq e_{\max}\}$ . The quantity  $u = \frac{1}{2}\beta^{1-t}$  is referred to as the *relative precision* of floating point number system  $\mathcal{F}$ .

The *range* of  $\mathcal{F}$  is the set  $\mathcal{G} := \{x \in \mathbb{R} \mid \beta^{e_{\min}-1} \leq |x| \leq \beta^{e_{\max}}(1 - \beta^{-1})\}$ . Note that  $\mathcal{F} \subset \mathcal{G}$ . Consider the *rounding function*  $\text{fl} : \mathbb{R} \rightarrow \mathcal{F}, x \mapsto \arg\min_{y \in \mathcal{F}} |x - y|$ . One can show that every  $x \in \mathcal{G}$  satisfies  $\text{fl}(x) = x(1 + \delta) \in \mathcal{F}$  for some  $\delta$  with  $|\delta| \leq u$ . This is a crucial property. It tells us that every number in  $\mathcal{G}$  can be approximated by a number of  $\mathcal{F}$  with relative precision  $u$ . Namely, the following inequality holds:

$$\text{RelError}(x, \text{fl}(x)) = \|\delta\| \leq u \quad \text{for all } x \in \mathcal{G}. \tag{9.1}$$

Many hardware floating-point units use the IEEE 754 standard. This is a system  $\mathcal{F}$  with the specifications

|                 | $\beta$ | $t$ | $e_{\min}$ | $e_{\max}$      | $u$                       |
|-----------------|---------|-----|------------|-----------------|---------------------------|
| half (16 bit)   | 2       | 11  | -14        | $16 = 2^4$      | $\approx 5 \cdot 10^{-4}$ |
| single (32 bit) | 2       | 24  | -125       | $128 = 2^7$     | $\approx 6 \cdot 10^{-8}$ |
| double (64 bit) | 2       | 53  | -1021      | $1024 = 2^{10}$ | $\approx 10^{-16}$        |

This is designed so that the arithmetic operations  $\circ \in \{e+, -, \times, /, \sqrt{\cdot}\}$  satisfy  $\text{fl}(x \circ y) = (x \circ y)(1 + \delta)$  for some  $|\delta| \leq u$ . For instance, the 64-bit floating point number system, specified in the third row of the table, can approximate any real number within its range with a relative error of at most  $u \approx 10^{-16}$ .

After this digression to the practical aspects of numerical computing, we now return to the mathematical theory. The next theorem is also due to Rice [156]. In his article,  $M$  and  $N$  are arbitrary Riemannian manifolds. We here specialize to the algebraic setting. For us, each of  $M$  and  $N$  is a submanifold in a Euclidean space, described by a finite Boolean combination of polynomial equations and inequalities.

**Theorem 9.6** *Fix differentiable function  $\phi : M \rightarrow N$ . The condition number of this computational problem at  $\mathbf{x} \in M$  is the maximal norm of the derivative over the unit sphere in the tangent space at  $\mathbf{x}$ , i.e.*

$$\kappa[\phi](\mathbf{x}) = \max_{\mathbf{t} \in T_{\mathbf{x}}M : \|\mathbf{t}\|=1} \|D_{\mathbf{x}}\phi(\mathbf{t})\|.$$

We obtain the relative condition number  $\kappa_{\text{REL}}[\phi](\mathbf{x})$  by multiplying this quantity with  $\|\mathbf{x}\|/\|\phi(\mathbf{x})\|$ .

A key step in computing the condition number with Rice' formula is to find an epression for the Jacobian  $D_{\mathbf{x}}\phi$ . For problems of interest to us, this step uses implicit differentiation or geometric arguments.

**Example 9.7 (Roots of univariate polynomials)** Following [36, Section 14.1.1], we examine the computational problem of finding one real root of a polynomial  $f$  in one variable  $z$  of degree  $d$ . We write

$$f(z) = f_0 + f_1 z + f_2 z^2 + \cdots + f_d z^d.$$

The coefficient vector  $f = (f_0, f_1, \dots, f_d)$  now serves as the input  $\mathbf{x}$ , and the output is a particular a real number  $a$  which satisfies  $f(a) = 0$ . The function  $f \mapsto a(f)$  is well defined in a small open subset of the coefficient space  $\mathbb{R}^{d+1}$ . This must be small enough so that one root can be identified for each polynomial.

To find the derivative  $D_f a$  of our root-finding function  $f \mapsto a(f)$ , we assume that the coefficients are differentiable functions  $f_i(t)$  of a parameter  $t$ , and we set  $\dot{f}_i = f'_i(0)$  and  $\dot{f} = \sum_{i=0}^d i z^i$ . By differentiating the identity  $f(a(f)) = 0$  with respect to  $t$ , we find the following formula for the desired derivative:

$$D_f a = -\frac{\dot{f}(a)}{f'(a)}. \quad (9.2)$$

We now think of  $a$  as a fixed root of a fixed polynomial  $f(z)$ . Theorem 9.6 implies that the condition number  $\kappa[a](f)$  equals  $|f'(a)|^{-1}$  times the maximal value  $|\dot{f}(a)|$ , where  $\dot{f}$  runs over all points  $(\dot{f}_0, \dot{f}_1, \dots, \dot{f}_d)$  on the unit  $d$ -sphere. The Cauchy-Schwarz inequality implies that this maximum equals  $|\dot{f}(a)| = \sqrt{\sum_{i=0}^d a^{2i}}$ , and therefore we have the following formula for the condition number

$$\kappa[a](f) = \frac{\sqrt{\sum_{i=0}^d a^{2i}}}{f'(a)}. \quad (9.3)$$

This quantity is  $+\infty$  when  $a$  is a double zero of  $f$ . The further away from being a double zero, the smaller the condition number. The relative condition number for the univariate root finding problem equals

$$\kappa_{\text{REL}}[a](f) = \frac{\sqrt{(\sum_{i=0}^d f_i^2)(\sum_{i=0}^d a^{2i})}}{|a| |f'(a)|}.$$

We emphasize that this formula involves not just the polynomial  $f(z)$  but also the zero  $a$  we seek to find. This is due to the fact that  $a(f)$  is a *local map*, defined locally around the zero  $a$ . We also refer to [98].  $\diamond$

**Example 9.8 (The condition number of the ED minimization problem)** We view the ED minimization problem for a smooth algebraic variety  $X \subset \mathbb{R}^n$  from the point of view of condition numbers. Recall from Chapter 7 the definition of the medial axis  $\text{Med}(X)$ . Every point outside the medial axis has a unique closest point on  $X$ . Formulating this as a computational problem we have the input space  $M = \mathbb{R}^n \setminus \text{Med}(X)$  and the output space  $N = X$ . Here,  $\phi : (\mathbb{R}^n \setminus \text{Med}(X)) \rightarrow X$  is the function that projects  $\mathbf{u}$  to the closest point  $\phi(\mathbf{u}) = \operatorname{argmin}_{\mathbf{x} \in X} \|\mathbf{x} - \mathbf{u}\|$  on  $X$ . We can apply Theorem 9.6 to compute the associated condition number  $\kappa[\phi](\mathbf{u})$ . This was carried out in detail in [30]. Suppose  $\mathbf{x} = \phi(\mathbf{u})$ . We have

$$\kappa[\phi](\mathbf{u}) = \max_{\mathbf{t} \in T_{\mathbf{x}} X: \|\mathbf{t}\|=1} \| (I_m - \lambda \cdot L_{\mathbf{v}})^{-1} \mathbf{t} \|,$$

where  $\lambda = \|\mathbf{u} - \mathbf{x}\|$  is the distance from  $\mathbf{u}$  to  $X$ ,  $\mathbf{v} = \lambda^{-1}(\mathbf{u} - \mathbf{x})$  is the normal vector at  $\mathbf{x}$  pointing towards  $\mathbf{u}$ , and  $L_{\mathbf{v}}$  is the Weingarten map (6.8) of  $X$  at  $\mathbf{x}$  in normal direction  $\mathbf{v}$ ; see [30, Theorem 4.3].  $\diamond$

## 9.2 Matrix Inversion and Eckhart-Young

In this section we study the condition number for the problem of matrix inversion. This will lead us back to the Eckhart-Young theorem, for which we here present a proof. We begin by reviewing some norms on the space of real  $m \times n$  matrices. The usual Euclidean norm, or Frobenius norm, of a matrix  $A \in \mathbb{R}^{m \times n}$  is

$$\|A\| = \sqrt{\sum_{i,j} a_{ij}^2} = \sqrt{\operatorname{trace}(AA^T)}.$$

By contrast, in Theorem 9.6 we used the operator norm, or nuclear norm. This is defined as follows:

$$\|A\|_{\text{op}} := \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|.$$

If  $U$  and  $V$  are orthogonal matrices, then  $\|UAV^T\|_{\text{op}} = \|A\|_{\text{op}}$  and  $\|UAV^T\| = \|A\|$ . In words, *orthogonal invariance* holds for both norms. Suppose  $n \leq m$ . Let  $A = UDV^T$  be the singular value decomposition of  $A$ , where  $U \in \mathbb{R}^{m \times m}$  and  $V \in \mathbb{R}^{n \times n}$  are orthogonal matrices and  $D = \operatorname{diag}(\sigma_1, \dots, \sigma_n)$  is  $m \times n$  matrix with singular values  $\sigma_1 \geq \dots \geq \sigma_n \geq 0$  on the main diagonal. Orthogonal invariance implies  $\|A\|_{\text{op}} = \sigma_1$  and  $\|A\| = \sqrt{\sigma_1^2 + \dots + \sigma_n^2}$ . If  $n = m$  and  $\sigma_1 \neq 0$ , then we have  $A^{-1} = VD^{-1}U^T$ , and hence

$$\|A^{-1}\|_{\text{op}} = \sigma_1^{-1}.$$

We now focus on the case of square matrices ( $m = n$ ). Matrix inversion corresponds to the map

$$\operatorname{inv} : \mathcal{D} \rightarrow \mathcal{D}, A \mapsto A^{-1},$$

where  $\mathcal{D} = \{A \in \mathbb{R}^{n \times n} \mid \det(A) \neq 0\}$ . We shall prove the following characterization of the condition number of matrix inversion. For any  $A \in \mathcal{D}$ , the smallest singular value  $\sigma_n$  is a positive real number.

**Theorem 9.9** *The condition number of matrix inversion at  $A \in \mathcal{D}$  is*

$$\kappa[\text{inv}](A) = \|A^{-1}\|_{\text{op}}^2 = \sigma_n^{-2}.$$

Before we prove this theorem, let us briefly bring it in context with Remark 9.5. In numerical analysis, a popular choice for measuring the relative error is using the operator norm. In this case, by Theorem 9.9 the relative condition number can be expressed as the ratio of the largest and smallest singular value of  $A$ :

$$\kappa_{\text{REL}}[\text{inv}](A) = \kappa[\text{inv}](A) \cdot \frac{\|A\|_{\text{op}}}{\|A^{-1}\|_{\text{op}}} = \frac{\sigma_1}{\sigma_n}$$

This ratio is known as *Turing's condition number* and goes back to the work of Turing [179].

**Proof (of Theorem 9.9)** Let  $\text{adj}(A)$  denote the adjoint matrix of  $A$ . Since  $\text{inv}(A) = A^{-1} = \frac{1}{\det(A)} \cdot \text{adj}(A)$ , the map  $\text{inv}$  is a polynomial function on  $\mathcal{D}$ . In particular,  $\text{inv}$  is differentiable on  $\mathcal{D}$ . By Theorem 9.6 the condition number is  $\kappa[\text{inv}](A) = \|D_A \text{inv}\|_{\text{op}}$ . We compute the derivative of  $\text{inv}$ . Taking the derivative of  $AB = \mathbf{I}_n$  we have  $\dot{A}B + A\dot{B} = 0$ . Since  $B = A^{-1}$ , we conclude  $\dot{B} = -A^{-1}\dot{A}A^{-1}$ . For a tangent vector  $\dot{A} = R \in \mathbb{R}^{n \times n}$  we therefore have  $D_A \text{inv}(R) = A^{-1}RA^{-1}$ . Let  $A = UDV^T$  be the singular value decomposition of  $A$ . Then,  $A^{-1} = VD^{-1}U^T$ . By orthogonal invariance of the Euclidean norm, we find

$$\max_{\|R\|=1} \|A^{-1}RA^{-1}\|^2 = \max_{\|R\|=1} \|D^{-1}RD^{-1}\|^2 = \max_{\sum_{i,j} r_{i,j}^2=1} \sum_{i,j} \frac{r_{i,j}^2}{\sigma_i^2 \sigma_j^2}.$$

The last expression is maximized when  $r_{n,n} = 1$  and all other  $r_{i,j}$  are zero. We conclude that the condition number of matrix inversion at  $A$  is equal to  $\kappa[\text{inv}](A) = \max_{\|R\|=1} \|D_A \text{inv}(R)\|_{\text{op}} = \frac{1}{\sigma_n^2} = \|A^{-1}\|_{\text{op}}^2$ .  $\square$

**Example 9.10** We revisit Example 9.1. The two matrices were  $A = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$  and  $B = \begin{pmatrix} 1 & 1 \\ 1 & 1+\delta \end{pmatrix}$ . To be concrete, we set  $\delta = 10^{-8}$ . Since  $A^T A = 2 \cdot \mathbf{1}_2$ , we have  $\|A^{-1}\mathbf{x}\| = \frac{1}{2}\|\mathbf{x}\|$  for all  $\mathbf{x} \in \mathbb{R}^2$ . By Theorem 9.9,

$$\kappa[\text{inv}](A) = \|A^{-1}\|_{\text{op}}^2 = \frac{1}{4}.$$

On the other hand, and  $\|B^{-1}\|_{\text{op}} \geq \|B^{-1}\mathbf{e}_0\| \geq 10^8$ , so that

$$\kappa[\text{inv}](B) = \|B^{-1}\|_{\text{op}}^2 \geq 10^{16}.$$

This explains the different behaviors of the outputs with respect to errors in the input in Example 9.1.  $\diamond$

The Eckhart-Young Theorem (Theorem 2.10) yields a metric interpretation of Turing's condition number  $\kappa[\text{inv}](A)$  from Theorem 9.9. The smallest singular value  $\sigma_n$  of a square matrix  $A$  equals the Euclidean distance of  $A$  to the variety of singular matrices. This is the hypersurface defined by the determinant:

$$\Sigma := \{A \in \mathbb{R}^{n \times n} \mid \det(A) = 0\}.$$

Thus, Turing's condition number of the matrix  $A$  is the inverse distance (squared) to the hypersurface  $\Sigma$ :

$$\kappa[\text{inv}](A) = \frac{1}{\text{dist}(A, \Sigma)^2} \quad \text{and} \quad \kappa_{\text{REL}}[\text{inv}](A) = \frac{\|A\|_{\text{op}}}{\text{dist}(A, \Sigma)}. \quad (9.4)$$

**Example 9.11** Fix the  $2 \times 2$ -determinant  $\Sigma = V(\det) \subset \mathbb{R}^{2 \times 2}$  and consider the set of matrices  $A \in \mathbb{R}^{2 \times 2}$  for which  $\kappa[\text{inv}](A) = \varepsilon^{-1}$ . This is the offset hypersurface of  $\Sigma$  at level  $\sqrt{\varepsilon}$ . We can compute this as in Example 7.12. In what follows we compute the hypersurface defined by  $\kappa_{\text{REL}}[\text{inv}](A) = \varepsilon^{-1}$  for  $\varepsilon > 0$ .

To make the formula in (9.4) more convenient, we replace  $\|A\|_{\text{op}}$  by  $\|A\|$ . We proceed as in Section 7.2 to compute a polynomial equation for  $\text{dist}(A, \Sigma) = \varepsilon \cdot \|A\|$  in terms of  $A$  and  $\varepsilon$ , but for  $B \in \Sigma$  we replace the affine sphere  $\|A - B\| = \varepsilon$  by the homogeneous sphere  $\|A - B\| = \varepsilon \cdot \|A\|$ . We use **Macaulay2** [79]:

```
R = QQ[a_0..a_3, b_0..b_3, eps];
f = b_0*b_3 - b_1*b_2;
normAsq = a_0^2 + a_1^2 + a_2^2 + a_3^2;
d = (a_0-b_0)^2 + (a_1-b_1)^2 + (a_2-b_2)^2 + (a_3-b_3)^2 - eps^2 * normAsq;
J1 = {diff(b_0, f), diff(b_1, f), diff(b_2, f), diff(b_3, f)};
J2 = {diff(b_0, d), diff(b_1, d), diff(b_2, d), diff(b_3, d)};
J = matrix {J1, J2};
OC = ideal {f, minors(2, J), d};
O = eliminate({{b_0, b_1, b_2, b_3}}, OC);
g = (gens O)_0_0
```

The result equals  $g(\mathbf{a}, \varepsilon) = (a_0^2 + a_1^2 + a_2^2 + a_3^2)^2 \cdot (g_0(\mathbf{a}) + g_1(\mathbf{a})\varepsilon^2 + g_2(\mathbf{a})\varepsilon^4 + g_3(\mathbf{a})\varepsilon^6 + g_4(\mathbf{a})\varepsilon^8)$ , where

$$g_4(\mathbf{a}) = (a_0^2 + a_1^2 + a_2^2 + a_3^2)^2 \quad (9.5)$$

$$g_3(\mathbf{a}) = -3(a_0^2 + a_1^2 + a_2^2 + a_3^2)^2 \quad (9.6)$$

$$g_2(\mathbf{a}) = 3a_0^4 + 6a_0^2a_1^2 + 3a_1^4 + 6a_0^2a_2^2 + 7a_1^2a_2^2 + 3a_2^4 - 2a_0a_1a_2a_3 + 7a_0^2a_3^2 + 6a_1^2a_3^2 + 6a_2^2a_3^2 + 3a_3^4 \quad (9.7)$$

$$g_1(\mathbf{a}) = -(a_0^4 + 2a_0^2a_1^2 + a_1^4 + 2a_0^2a_2^2 + 4a_1^2a_2^2 + a_2^4 - 4a_0a_1a_2a_3 + 4a_0^2a_3^2 + 2a_1^2a_3^2 + 2a_2^2a_3^2 + a_3^4) \quad (9.8)$$

$$g_0(\mathbf{a}) = (a_1a_2 - a_0a_3)^2. \quad (9.9)$$

Figure 9.1 shows the zero set of  $g(\mathbf{a}, \varepsilon)$  at level  $\varepsilon = 0.5$  in the affine patch  $a_0 = 1$ . We remark that, if  $\text{dist}(A, \Sigma) = \varepsilon \cdot \|A\|$ , then  $\varepsilon = \sin \alpha$ , where  $\alpha$  is the minimal angle between the line  $\mathbb{R} \cdot A$  and a line  $\mathbb{R} \cdot B$  with  $B \in \Sigma$ . In this case,  $\varepsilon^{-1}$  is also called a *conic condition number* (see [36, Chapters 20 & 21]).  $\diamond$

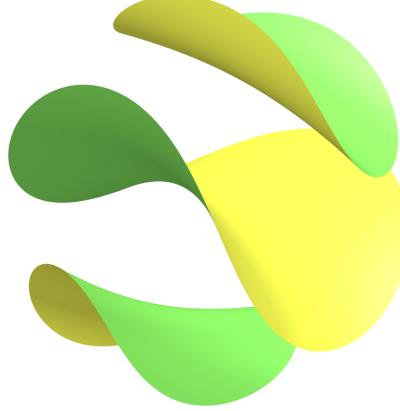


Fig. 9.1: The picture shows the determinantal hypersurface  $\Sigma = \{\det(A) = 0\} \subset \mathbb{R}^{2 \times 2}$  (the surface in the middle) together with the hypersurface defined by  $\text{dist}(A, \Sigma) = \varepsilon \cdot \|A\|$  with  $\varepsilon = 0.5$  (the union of the two surfaces on the outside) in the affine patch where the upper left entry of  $A$  is fixed to be one. The image was created using **Surfer**.

In the remainder of this section we give a proof of the Eckart-Young theorem. For this we return now to rectangular matrices and denote by  $X_r := \{A \in \mathbb{R}^{m \times n} \mid \text{rank}(A) \leq r\}$  the variety of real matrices of rank at most  $r$ . Recall that  $\text{Sing}(X_r) = X_{r-1}$ . We first compute the normal space of  $X_r$  at a smooth point. Lemma 9.12 can be viewed as a variant of Example 2.22, but presented in a more down-to-earth manner.

**Lemma 9.12** Let  $A \in X_r$  be a smooth point (i.e.,  $\text{rank}(A) = r$ ). Suppose that  $A = RS^T$ , where  $R \in \mathbb{R}^{m \times r}$  and  $S \in \mathbb{R}^{n \times r}$  have rank  $r$ . The normal space of  $X_r$  at  $A$  has dimension  $(m-r)(n-r)$  and it equals

$$N_A X_r = \text{span} \{ \mathbf{u}\mathbf{v}^T \mid \mathbf{u}^T R = 0 \text{ and } S^T \mathbf{v} = 0 \}.$$

**Proof** Let  $R(t) \in \mathbb{R}^{m \times r}$  and  $S(t) \in \mathbb{R}^{n \times r}$  be smooth curves with  $R(0) = R$  and  $S(0) = S$ . Then,  $\gamma(t) := R(t)S(t)^T$  is a smooth curve with  $\gamma(0) = A$ . The product rule from calculus gives

$$\frac{\partial}{\partial t} \gamma(t) \Big|_{t=0} = R \left( \frac{\partial}{\partial t} S(t) \Big|_{t=0} \right)^T + \left( \frac{\partial}{\partial t} R(t) \Big|_{t=0} \right) S^T. \quad (9.10)$$

Let  $\mathcal{V} := \{RP^T \mid P \in \mathbb{R}^{n \times r}\}$  and  $\mathcal{W} := \{QS^T \mid Q \in \mathbb{R}^{m \times r}\}$ . The equation (9.10) shows that the tangent space of  $X_r$  at  $A$  is the sum  $T_A X_r = \mathcal{V} + \mathcal{W}$ . (Aside for students: what is the intersection  $\mathcal{V} \cap \mathcal{W}$ ?)

We note that  $\mathcal{V}$  consists of all matrices  $L \in \mathbb{R}^{m \times n}$  such that  $\mathbf{u}^T L \mathbf{x} = 0$  for all  $\mathbf{u}$  with  $\mathbf{u}^T R = 0$  and  $\mathbf{x} \in \mathbb{R}^n$  arbitrary. Since  $\mathbf{u}^T L \mathbf{x} = \text{Trace}(L^T \mathbf{u} \mathbf{x}^T)$ , this shows that the normal space of  $\mathcal{V}$  is spanned by matrices of the form  $\mathbf{u}\mathbf{x}^T$ . Similarly, the normal space of  $\mathcal{W}$  is spanned by  $\mathbf{y}\mathbf{v}^T$ , where  $S^T \mathbf{v} = 0$  and  $\mathbf{y} \in \mathbb{R}^m$  arbitrary. Therefore, the normal space of  $T_A X_r = \mathcal{V} + \mathcal{W}$  is spanned by all  $\mathbf{u}\mathbf{v}^T$  with  $\mathbf{u}$  and  $\mathbf{v}$  as above.  $\square$

**Corollary 9.13**  $\dim X_r = nm - (m-r)(n-r) = r(m+n-r)$ .

We use Lemma 9.12 to prove the Eckart-Young Theorem.

**Proof (of Theorem 2.10)** Let  $B \in X_r$  be a matrix of rank  $r$ . We consider a singular value decomposition  $B = UDV^T$ . The matrices  $U \in \mathbb{R}^{m \times r}$  and  $V \in \mathbb{R}^{r \times n}$  have orthonormal columns, and  $D = \text{diag}(\sigma_1, \dots, \sigma_r)$  with  $\sigma_1, \dots, \sigma_r > 0$  (not necessarily ordered). To derive Theorem 2.10, we shall prove a claim that is reminiscent of Theorem 8.21. Namely, the singular value decomposition of all matrices  $A \in \mathbb{R}^{m \times n}$ , such that  $B$  is an ED-critical point for  $A$ , has the form  $A = [U' U'] \begin{pmatrix} D & 0 \\ 0 & D' \end{pmatrix} [V' V]^T$ , where  $D' = \text{diag}(\sigma_{r+1}, \dots, \sigma_n)$ .

By orthogonal invariance, we can assume that  $B = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}$ . Let  $A \in B + N_B X_r$  be a matrix in the normal space of  $B$ . By Lemma 9.12, we have  $A = B + \sum_{i=r+1}^m \sum_{j=r+1}^n a_{ij} \mathbf{e}_i \mathbf{e}_j^T$  for some coefficients  $a_{ij} \in \mathbb{R}$ , i.e.

$$A = \begin{pmatrix} D & 0 \\ 0 & A' \end{pmatrix}, \quad \text{where } A' = (a_{ij}) \in \mathbb{R}^{(m-r) \times (n-r)}.$$

Let now  $A' = U'D'V'$  be the singular value decomposition of  $A'$ . Then,

$$A = [I_r \ U'] \begin{pmatrix} D & 0 \\ 0 & D' \end{pmatrix} [I_r \ V']^T. \quad (9.11)$$

This must be the desired singular value decomposition of the  $m \times n$  matrix  $A$ , because the singular value decomposition of any rectangular matrix is unique up to ordering the singular values.  $\square$

**Remark 9.14** Our proof of the Eckart-Young Theorem also works for symmetric  $n \times n$  matrices. It implies that the rank  $r$  matrix  $B$  which minimizes the distance to a symmetric matrix  $A$  is also symmetric.

### 9.3 Condition Number Theorems

The formula in (9.4) states that the condition number of matrix inversion at the input  $A \in \mathbb{R}^{n \times n}$  is the inverse squared distance to the variety of singular matrices  $\Sigma$ . To a numerical analyst, the elements of a set like  $\Sigma$  are the *ill-posed inputs* of the computational problem. A result that connects a condition number

and an inverse distance to ill-posed inputs, is called a *condition number theorem*. Equation (9.4) yields a condition number theorem for matrix inversion. Condition number theorems were derived in [181] for computing eigenvalues of matrices, and in [98] for computing zeros of polynomials, as in Example 9.7.

Condition number theorems connect metric algebraic geometry with numerical analysis. They relate the numerical difficulty of an input to a problem to the distance of this input to the locus of ill-posed inputs. This is explained in detail in Demmel's paper [57]. The determinant of a square matrix is a special case of a discriminant, and  $\Sigma$  will generally be a discriminantal hypersurface. This will made precise in the next section. In this section, we prove a condition number theorem for solving systems of polynomial equations. We shall consider homogeneous polynomials and their zeros in projective space. This differs from Example 9.7, where  $a$  was a zero on the affine line. While this difference may seem insignificant to an algebraic geometer, it becomes significant in our metric setting. Namely, affine space with its Euclidean metric and projective space with the metric described below, are markedly different as metric spaces.

We fix the real projective space  $\mathbb{P}_{\mathbb{R}}^n$  of dimension  $n$ , and we write  $\mathcal{H}_d$  for the vector space of homogeneous polynomials of degree  $d$  in  $\mathbf{x} = (x_0, \dots, x_n)$ . For  $m \leq n$  and a tuple  $\mathbf{d} = (d_1, \dots, d_m)$  we abbreviate

$$\mathcal{H}_{\mathbf{d}} := \mathcal{H}_{d_1} \times \cdots \times \mathcal{H}_{d_m}.$$

Suppose  $m = n$ , let  $F \in \mathcal{H}_{\mathbf{d}}$  and let  $\mathbf{a} \in \mathbb{P}_{\mathbb{R}}^n$  a regular zero of  $F$ . By the Implicit Function Theorem, there is a locally defined solution function  $a : U \rightarrow V$ , where  $U \subset \mathcal{H}_{\mathbf{d}}$  is a neighborhood of  $F$  and  $V \subset \mathbb{P}_{\mathbb{R}}^n$  is a neighborhood of  $\mathbf{a}$ , such that  $F(a(F)) = 0$ . Our goal is to give a formula for the condition number of  $a$ .

For this, we first need to introduce metrics on the space of polynomials and the space of zeros. Since  $a$  is defined locally, we can replace projective space  $\mathbb{P}_{\mathbb{R}}^n$  by the  $n$ -sphere  $\mathbb{S}^n \subset \mathbb{R}^{n+1}$ . The latter carries the natural metric inherited from its ambient space. In fact, we can use the Riemannian metric on the sphere to define a metric on  $\mathbb{P}_{\mathbb{R}}^n$ , so that it becomes a Riemannian manifold (this approach is worked out in Section 12.3 for complex projective space). In this way,  $\mathbb{P}_{\mathbb{R}}^n$  becomes curved, while affine space is flat.

For  $\mathcal{H}_d$  we use the Bombieri-Weyl metric. Write  $A := \{\alpha \in \mathbb{N}^{n+1} \mid \alpha_0 + \cdots + \alpha_n = d\}$ . The *Bombieri-Weyl inner product* between two polynomials  $f = \sum_{\alpha \in A} f_{\alpha} \mathbf{x}^{\alpha} \in \mathcal{H}_d$  and  $g = \sum_{\alpha \in A} g_{\alpha} \mathbf{x}^{\alpha} \in \mathcal{H}_d$  is defined by

$$\langle f, g \rangle_{\text{BW}} := \sum_{\alpha \in A} \frac{\alpha_0! \cdots \alpha_n!}{d!} f_{\alpha} \cdot g_{\alpha}. \quad (9.12)$$

For  $d = 1$  this is the usual Euclidean inner product in  $\mathbb{R}^{n+1}$ . The reason for the multinomial coefficients in (9.12) is that this inner product is orthogonally invariant: if  $U \in O(n+1)$  is an orthogonal matrix, then

$$\langle f \circ U, g \circ U \rangle_{\text{BW}} = \langle f, g \rangle_{\text{BW}}.$$

Kostlan [115, 116] showed that the Bombieri-Weyl inner product (9.12) is the unique (up to scaling) orthogonally invariant inner product on  $\mathcal{H}_d$  such that monomials are pairwise orthogonal. This inner product extends to  $\mathcal{H}_{\mathbf{d}}$ . Namely, for tuples of polynomials  $F = (f_1, \dots, f_m)$  and  $G = (g_1, \dots, g_m)$ , we set

$$\langle F, G \rangle_{\text{BW}} := \langle f_1, g_1 \rangle_{\text{BW}} + \cdots + \langle f_m, g_m \rangle_{\text{BW}}.$$

The Bombieri-Weyl norm is  $\|F\|_{\text{BW}} := \sqrt{\langle F, F \rangle_{\text{BW}}}$ , and the distance corresponding to this norm is

$$\text{dist}_{\text{BW}}(F, G) := \|F - G\|_{\text{BW}} \quad \text{for } F, G \in \mathcal{H}_{\mathbf{d}}.$$

**Example 9.15** ( $n = 1, d = 2$ ) Consider two binary quadrics  $f(x_0, x_1) = ax_0^2 + bx_0x_1 + cx_1^2 = \mathbf{x}^T A \mathbf{x}$  and  $g(x_0, x_1) = \alpha x_0^2 + \beta x_0x_1 + \gamma x_1^2 = \mathbf{x}^T B \mathbf{x}$ , where  $\mathbf{x} = (x_0, x_1)^T$ , with corresponding symmetric matrices

$$A = \begin{pmatrix} a & b/2 \\ b/2 & c \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} \alpha & \beta/2 \\ \beta/2 & \gamma \end{pmatrix}.$$

The inner product of the two quadratic forms is the trace inner product of the two matrices:

$$\langle f, g \rangle_{\text{BW}} = \frac{2! \cdot 0!}{2!} a\alpha + \frac{1! \cdot 1!}{2!} b\beta + \frac{2! \cdot 0!}{2!} c\gamma = a\alpha + \frac{1}{2} b\beta + c\gamma = \text{Trace}(A^T B).$$

An analogous formula holds for quadrics in more than two variables. Thus, the Bombieri-Weyl product generalizes the trace inner product for symmetric matrices to homogeneous polynomials of any degree.  $\diamond$

We now turn to the condition number of the solution function  $a : U \rightarrow V$ . The zero  $\mathbf{a} = a(F)$  is represented by a unit vector in  $\mathbb{S}^n$ . Proceeding as in Example 9.7, we differentiate the equation  $F(a(F)) = 0$ . This gives  $JF(\mathbf{a})\dot{\mathbf{a}} + \dot{F}(\mathbf{a})$ . Here,  $JF(\mathbf{a}) = (\frac{\partial f_i}{\partial x_j}(\mathbf{a})) \in \mathbb{R}^{n \times (n+1)}$  is the Jacobian matrix of  $F$  at  $\mathbf{a}$ , and  $\dot{\mathbf{a}} \in T_{\mathbf{a}}\mathbb{S}^n$ . The tangent space  $T_{\mathbf{a}}\mathbb{S}^n$  is the image of the projection matrix  $P_{\mathbf{a}} := \text{Id}_{n+1} - \mathbf{a}\mathbf{a}^T$ . Note that  $\det(P_{\mathbf{a}}) = 1 - \|\mathbf{a}\|^2 = 0$ . When  $\mathbf{a}$  is a regular zero of  $F$ , the rank of  $JF(\mathbf{a})$  is  $n$ , and we obtain

$$\dot{\mathbf{a}} = (JF(\mathbf{a})|_{T_{\mathbf{a}}\mathbb{S}^n})^{-1} \dot{F}(\mathbf{a}).$$

This formula is a multivariate version of (9.2). By Theorem 9.6, the condition number for  $a$  is  $\kappa[a](F) = \max_{\|F\|_{\text{BW}}=1} \|(JF(\mathbf{a})|_{T_{\mathbf{a}}\mathbb{S}^n})^{-1} \dot{F}(\mathbf{a})\|$ . We use orthogonal invariance of the Bombieri-Weyl metric to compute this maximum. Fix  $U \in O(n+1)$  such that  $\mathbf{a} = U\mathbf{e}_0$ , where  $\mathbf{e}_0 = (1, 0, \dots, 0) \in \mathbb{S}^n$ . We have  $\|\dot{F}\|_{\text{BW}} = \|\dot{F} \circ U\|_{\text{BW}}$  for all  $\dot{F} \in \mathcal{H}_{\mathbf{d}}$ . This implies that  $\kappa[a](F)$  is also obtained by maximizing  $(JF(\mathbf{a})|_{T_{\mathbf{a}}\mathbb{S}^n})^{-1} \dot{F}(\mathbf{e}_0)$  over all  $\dot{F} \in \mathcal{H}_{\mathbf{d}}$  with  $\|F\|_{\text{BW}} = 1$ . Let us write  $\dot{F}(\mathbf{x}) = x_0^d \cdot \mathbf{b} + G(\mathbf{x})$ , where  $\mathbf{b} \in \mathbb{R}^n$  and  $G$  involves only powers of  $x_0$  of degree less than  $d$ . Then,  $\dot{F}(\mathbf{e}_0) = \mathbf{b}$  and so  $\|\dot{F}(\mathbf{e}_0)\| = \|\mathbf{b}\| = \sqrt{1 - \|G\|_{\text{BW}}^2}$ . This uses the fact that  $x_0^d \cdot \mathbf{b}$  and  $G$  are orthogonal for the Bombieri-Weyl inner product. This shows that  $\{\dot{F}(\mathbf{a}) : \|\dot{F}\|_{\text{BW}} = 1\}$  is the unit ball in  $\mathbb{R}^{n+1}$ . Consequently,

$$\kappa[a](F) = \|(JF(\mathbf{a})|_{T_{\mathbf{a}}\mathbb{S}^n})^{-1}\|_{\text{op}}. \quad (9.13)$$

**Corollary 9.16** *The condition number for solving  $n$  homogeneous polynomial equations on  $\mathbb{P}^n$  equals*

$$\kappa[a](F) = \frac{1}{\sigma_n(JF(\mathbf{a}) P_{\mathbf{a}})}. \quad (9.14)$$

**Proof** The matrix  $JF(\mathbf{a})P_{\mathbf{a}}$  represents the linear map  $JF(\mathbf{a})|_{T_{\mathbf{a}}\mathbb{S}^n}$ , so they have the same singular values.  $\square$

**Example 9.17 (Univariate polynomials revisited)** We examine our formula for  $n = 1$ . The computation takes a binary form  $f(x_0, x_1)$  to one of its zeros  $\mathbf{a} = (a_0, a_1)$  in  $\mathbb{S}^1$ . This is the projective version of Example 9.7. The Jacobian equals  $Jf(\mathbf{a}) = [\partial f / \partial x_0(\mathbf{a}) \ \partial f / \partial x_1(\mathbf{a})] \in \mathbb{R}^{1 \times 2}$ . Equation (9.14) yields

$$\kappa[a](F) = \frac{1}{\sqrt{(\partial f / \partial x_0(\mathbf{a}))^2 + (\partial f / \partial x_1(\mathbf{a}))^2}}.$$

This formula is the analogue of (9.3). Here,  $\mathbf{a}$  is a point on the circle  $\mathbb{S}^1$ , representing a zero of  $f$  in  $\mathbb{P}^1$ .  $\diamond$

For the next result we return to the general case  $m \leq n$ , i.e. we allow systems with fewer equations.

**Proposition 9.18** *For  $\mathbf{x} \in \mathbb{P}_{\mathbb{R}}^n$  denote  $\Sigma(\mathbf{x}) := \{F \in \mathcal{H}_{\mathbf{d}} \mid F(\mathbf{x}) = 0 \text{ and } \text{rank } JF(\mathbf{x}) < m\}$ . Then,*

$$\text{dist}_{\text{BW}}(F, \Sigma(\mathbf{x})) = \sqrt{\|F(\mathbf{x})\|^2 + \sigma_m(D^{-1/2} JF(\mathbf{x}) P_{\mathbf{x}})^2},$$

where  $D = \text{diag}(d_1, \dots, d_m)$ . Hence, if  $\mathbf{a}$  is a zero of  $F$ , then  $\text{dist}_{\text{BW}}(F, \Sigma(\mathbf{a})) = \sigma_m(D^{-1/2} JF(\mathbf{a}) P_{\mathbf{a}})$ .

**Proof** Let  $U \in O(n+1)$  with  $\mathbf{x} = U\mathbf{e}_0$  and set  $F_0 := F \circ U$ . By orthogonal invariance,  $\text{dist}_{\text{BW}}(F, \Sigma_{\mathbb{R}}(\mathbf{x})) = \text{dist}_{\text{BW}}(F_0, \Sigma_{\mathbb{R}}(\mathbf{e}_0))$ . We shall compute the latter. We write  $F_0(\mathbf{x}) = x_0^d \cdot \mathbf{b} + x_0^{d-1} \cdot B(x_1, \dots, x_n)^T + H(\mathbf{x})$ , where  $\mathbf{b} \in \mathbb{R}^m$ ,  $B \in \mathbb{R}^{m \times n}$  and  $H$  has degree  $\leq d-2$  in  $x_0$ . Note that  $\mathbf{b} = F_0(\mathbf{e}_0) = F(\mathbf{x})$  and  $B = JF_0(\mathbf{e}_0) P_{\mathbf{e}_0} = JF(\mathbf{x}) P_{\mathbf{x}} U^T$ . Recall that  $G \in \Sigma(\mathbf{e}_0)$  if and only if  $G(\mathbf{e}_0) = 0$  and the Jacobian  $JG(\mathbf{e}_0)$  has rank at most  $m-1$ . This means that the distance from  $\Sigma(\mathbf{e}_0)$  to  $F_0$  is minimized at a polynomial system of the form  $G_0(\mathbf{x}) = x_0^{d-1} \cdot A(x_1, \dots, x_n)^T \in \Sigma(\mathbf{e}_0)$ , where  $A$  is a matrix of rank at most  $m-1$ . We have

$$\|F_0 - G_0\|_{\text{BW}}^2 = \|\mathbf{b}\|^2 + \|D^{-1/2}(A - B)\|^2.$$

The Eckhart-Young Theorem implies that the distance from  $D^{-1/2}B$  to the variety of matrices of rank at most  $m-1$  equals  $\sigma_m(D^{-1/2}B)$ . From this we infer the desired formula for the distance to the variety  $\Sigma(\mathbf{x})$ :

$$\text{dist}_{\text{BW}}(F, \Sigma(\mathbf{x})) = \sqrt{\|\mathbf{b}\|^2 + \sigma_m(D^{-1/2}B)^2} = \sqrt{\|F(\mathbf{x})\|^2 + \sigma_m(D^{-1/2}JF(\mathbf{x})P_{\mathbf{x}})^2}.$$

Here we use that the singular values are invariant under multiplication with the orthogonal matrix  $U^T$ .  $\square$

Proposition 9.18 can be viewed as condition number theorem. If  $m = n$  and  $\mathbf{a}$  is a zero of  $F$  then the formula for the condition number in (9.14) is equal to the reciprocal distance to  $\Sigma(\mathbf{a})$ , up to the scaling factor by the diagonal matrix  $D$ . The variety  $\Sigma(\mathbf{a})$  plays the role of a discriminant, but it still depends on the point  $\mathbf{a}$ . To get the actual discriminant, we take the union over all points  $\mathbf{a}$  in  $\mathbb{P}^n$ . Namely, we define

$$\Sigma := \{F \in \mathcal{H}_{\mathbf{d}} \mid \text{there is } \mathbf{x} \in \mathbb{P}_{\mathbb{R}}^n \text{ s.t. } F(\mathbf{x}) = 0 \text{ and } \text{rank } JF(\mathbf{x}) < m\}, \quad (9.15)$$

Note that  $\Sigma$  is the discriminant, since we are asking for  $\mathbf{x}$  to be a real point. In the next section we replace  $\Sigma$  by its Zariski closure. This is the variety which is obtained by allowing  $\mathbf{x}$  to be complex.

The next theorem gives a formula for the distance from a polynomial system  $F$  to the discriminant  $\Sigma$ . This was proved by Rafalli [155] for the case  $m = 1$ . Bürgisser and Cucker [36] cover the case  $m = n$ .

**Theorem 9.19** *Let  $m \leq n$  and  $\mathbf{d} = (d_1, \dots, d_m)$  be a tuple of degrees. Let  $F \in \mathcal{H}_{\mathbf{d}}$ . Then,*

$$\text{dist}_{\text{BW}}(F, \Sigma_{\mathbb{R}}) = \min_{\mathbf{x} \in \mathbb{S}^n} \sqrt{\|F(\mathbf{x})\|^2 + \sigma_m(D^{-1/2}JF(\mathbf{x})P_{\mathbf{x}})^2},$$

where, as before,  $D = \text{diag}(d_1, \dots, d_m)$  and  $P_{\mathbf{x}} := I_{n+1} - \mathbf{x}\mathbf{x}^T$  is the projection onto  $T_{\mathbf{x}}\mathbb{S}^n$ .

**Proof** By definition, the discriminant equals  $\Sigma = \bigcup_{\mathbf{x} \in \mathbb{S}^n} \Sigma(\mathbf{x})$ . This implies

$$\text{dist}_{\text{BW}}(F, \Sigma) = \min_{\mathbf{x} \in \mathbb{S}^n} \text{dist}_{\text{BW}}(F, \Sigma(\mathbf{x}))$$

The minimum is obtained because  $\mathbb{S}^n$  is compact. The statement now follows from Proposition 9.18.  $\square$

**Remark 9.20** Theorem 9.19 can be generalized as follows. Fix  $1 \leq k < m$ . The distance of  $F \in \mathcal{H}_{\mathbf{d}}$  to the space of polynomial systems  $F \in \mathcal{H}_{\mathbf{d}}$ , such that there exists  $\mathbf{x} \in \mathbb{P}^n$  with  $F(\mathbf{x}) = 0$  and  $\text{rank } JF(\mathbf{x}) < k$  is

$$\min_{\mathbf{x} \in \mathbb{S}^n} \sqrt{\|F(\mathbf{x})\|^2 + \sum_{i=k}^m \sigma_i(D^{-1/2}JF(\mathbf{x})P_{\mathbf{x}})^2},$$

where  $\sigma_k(\cdot), \dots, \sigma_m(\cdot)$  are the  $m-k+1$  smallest singular values. This is because the distance of a matrix  $A \in \mathbb{R}^{m \times n}$  to the nearest matrix of rank at most  $k-1$  is  $\sqrt{\sum_{i=k}^m \sigma_i(A)^2}$ , by the Eckart-Young Theorem.

**Remark 9.21** Let  $m = 1$ . We have only one polynomial  $f \in \mathcal{H}_d$ , and Theorem 9.19 yields  $\text{dist}_{\text{BW}}(f, \Sigma_{\mathbb{R}}) = \min_{\mathbf{x} \in \mathbb{S}^n} \sqrt{f(\mathbf{x})^2 + \frac{1}{d} \|\nabla f(\mathbf{x}) P_{\mathbf{x}}\|^2}$ , where the row vector  $\nabla f(\mathbf{x})$  is the gradient of  $f$  at  $\mathbf{x}$ . By Euler's formula for homogeneous functions,  $\nabla f(\mathbf{x}) \cdot \mathbf{x} = d \cdot f(\mathbf{x})$ , so  $P_{\mathbf{x}} \nabla f(\mathbf{x}) = \nabla f(\mathbf{x}) - (d \cdot f(\mathbf{x})) \mathbf{x}$ . We get

$$\text{dist}_{\text{BW}}(f, \Sigma) = \min_{\mathbf{x} \in \mathbb{S}^n} \sqrt{f(\mathbf{x})^2 + \frac{1}{d} \|\nabla f(\mathbf{x}) - (d \cdot f(\mathbf{x})) \mathbf{x}\|^2}. \quad (9.16)$$

**Example 9.22** ( $m = 1, n = 2, d = 3$ ) The discriminant  $\Sigma$  for ternary cubics can be computed as follows:

```
R = QQ[x,y,z,c0,c1,c2,c3,c4,c5,c6,c7,c8,c9];
c0*x^3+c1*y^3+c2*z^3+c3*x^2*y+c4*x^2*z+c5*x*y^2+c6*y^2*z+c7*x*z^2+c8*y*z^2+c9*x*y*z;
I = ideal(diff(x,oo),diff(y,oo),diff(z,oo), z-1);
disc = first first entries gens eliminate({x,y,z},I);
toString disc
degree disc, # terms disc
```

We see that the discriminant  $\Sigma$  is given by a polynomial with 2040 terms of degree 12 in 10 unknowns.

We now compute the distance to  $\Sigma$  from the Fermat cubic  $f(\mathbf{x}) = x_0^3 + x_1^3 + x_2^3$ . This is easy if we use (9.16). Let  $h(\mathbf{x}) = (x_0^3 + x_1^3 + x_2^3)^2 + 3 \sum_{i=0}^2 (x_i^2 - (x_0^3 + x_1^3 + x_2^3) x_i)^2$ . Then  $\text{dist}_{\text{BW}}(f, \Sigma_{\mathbb{R}}) = \min_{\mathbf{x} \in \mathbb{S}^n} \sqrt{h(\mathbf{x})}$ . The polynomial function  $h : \mathbb{S}^2 \rightarrow \mathbb{R}$  is minimized at the point  $\mathbf{x}_0 = \frac{1}{\sqrt{3}}(1, 1, 1)$ , and we conclude that

$$\text{dist}_{\text{BW}}(f, \Sigma) = \sqrt{h(\mathbf{x}_0)} = \frac{1}{\sqrt{3}}.$$

The norm of the Fermat Cubic is  $\|f\|_{\text{BW}} = \sqrt{3}$ . Since  $\Sigma$  is a cone in  $\mathcal{H}_3$ , the minimal angle (measured in the Bombieri-Weyl metric) between  $f$  and any polynomial in  $\Sigma$  therefore is  $\arcsin(1/3) \approx 0.21635 \cdot \frac{\pi}{2}$ .  $\diamond$

## 9.4 Distance to the Discriminant

Theorem 9.19 expresses the distance from a polynomial system  $F$  to its discriminant  $\Sigma$  as the optimal value of an optimization problem over the unit sphere  $\mathbb{S}^n$ . Thus we are solving the ED problem from Chapter 2 for discriminants. In this section we examine this problem through the lens of algebraic geometry. This uses the ED duality in Theorem 2.23 and the fact that discriminants are projectively dual to toric varieties.

In the previous section we considered dense systems, where the polynomials involve all monomials of a fixed degree. In that setting, we used the orthogonal invariance of the Bombieri-Weyl metric to prove Theorem 9.19. In many applications, however, polynomial systems are not dense but *sparse*, and there is no such invariant metric. To address this, we now work in the setting of  $n$  sparse Laurent polynomials in  $n$  variables. Our equations are not homogeneous. The given data is a collection of finite support sets  $\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_n \subset \mathbb{Z}^n$ . Our system of equations takes the form  $f_1(\mathbf{x}) = f_2(\mathbf{x}) = \dots = f_n(\mathbf{x}) = 0$ , where

$$f_i(\mathbf{x}) = \sum_{\mathbf{a} \in \mathcal{A}_i} c_{i,\mathbf{a}} \mathbf{x}^{\mathbf{a}} \quad \text{for } i = 1, 2, \dots, n. \quad (9.17)$$

The BKK Theorem [19, 20] tells us that the number of solutions in  $(\mathbb{C}^*)^n$  equals the mixed volume  $\text{MV}(P_1, P_2, \dots, P_n)$ . Here,  $P_i$  is the convex hull of  $\mathcal{A}_i$  and the coefficients  $c_{i,\mathbf{a}}$  are assumed to be generic. The BKK theorem was discussed in Example 3.31. We here assume that the mixed volume is at least 2.

The *discriminant* of the system (9.17) is the irreducible polynomial  $\Delta(\mathbf{c})$  which vanishes whenever our equations have a double root in  $(\mathbb{C}^*)^n$ . Here  $\mathbf{c} = (c_{i,\mathbf{a}})$  denotes the vector of all coefficients. The discriminant  $\Delta$  is unique up to scaling. We refer [44] for many details about  $\Delta$ , including a more precise

definition. Our goal here is to solve the Euclidean distance problem for the discriminant hypersurface

$$\Sigma = \{ \mathbf{c} \in \mathbb{R}^N \mid \Delta(\mathbf{c}) = 0 \}. \quad (9.18)$$

*Remark 9.23* The real discriminant equals the real locus in the discriminant  $\Sigma$ . This means that the real hypersurface defined (for the dense case) in (9.15) agrees with the definition above. Indeed, let  $\mathbf{c}$  be a generic real point  $\mathbf{c}$  in the hypersurface  $\Sigma$ . Then the corresponding polynomial system  $F = (f_1, \dots, f_n)$  has a complex double zero  $\mathbf{x}$ . That double zero  $\mathbf{x}$  is unique. Since  $F$  has real coefficients,  $\mathbf{x}$  is invariant under complex conjugation. Therefore  $\mathbf{x}$  has real coordinates, and  $\mathbf{c}$  is also in the set (9.15). In fact, there is a formula for expressing  $\mathbf{x}$  by rational operations in  $\mathbf{c}$ . See (9.19) and [76, eqn (1.28) in Section 12.1.B].

From now on, we work over  $\mathbb{C}$  and we view the discriminant  $\Sigma$  as a complex projective hypersurface, defined by an irreducible polynomial with integer coefficients. This is consistent with [76] and Chapter 2.

**Example 9.24** Let  $n = 1$  and  $\mathcal{A} = \{0, 2, 5, 6\}$ . The discriminant of  $f(x) = c_1 + c_2x^2 + c_3x^5 + c_4x^6$  equals

$$\Delta = 46656c_1^4c_4^5 + 32400c_1^3c_2c_3^2c_4^3 - 3125c_1^3c_3^6 + 13824c_1^2c_2^3c_4^4 - 1500c_1^2c_2^2c_3^4c_4 + 192c_1c_2^4c_3^2c_4^2 + 1024c_2^6c_4^3 - 108c_2^5c_3^4.$$

This defines a surface  $\Sigma$  of degree 9 in  $\mathbb{P}^3$ . For any point  $\mathbf{c} \in \Sigma$ , we find the double root  $x$  of  $f$  as follows:

$$\left( \frac{\partial \Delta}{\partial c_1} : \frac{\partial \Delta}{\partial c_2} : \frac{\partial \Delta}{\partial c_3} : \frac{\partial \Delta}{\partial c_4} \right)(\mathbf{c}) = (1 : x^2 : x^5 : x^6). \quad (9.19)$$

This follows from duality between discriminants and toric varieties. In particular, if  $\mathbf{c}$  is real then so is  $x$ .  $\diamond$

Given any sparse system  $F$ , we seek the distance from its coefficient vector  $\mathbf{u}$  to the discriminant  $\Sigma$ . We will use the formulation of the ED problem given in Theorem 2.23. Let  $X$  be the toric variety naturally associated with the tuple  $(\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_n)$ . Then the dual variety  $X^\vee$  is precisely the discriminant hypersurface  $\Sigma$ . This observation is known as the *Cayley trick*; see [76] and [44, Section 2] for expositions.

We write  $\mathbf{x}$  for a point in  $X$  and  $\mathbf{c}$  for a point in  $\Sigma$ . The conormal variety  $N_X$  consists of pairs  $(\mathbf{x}, \mathbf{c})$  such that the hyperplane  $\mathbf{c}$  is tangent to  $X$  at the point  $\mathbf{x}$ . Equivalently,  $\mathbf{c}$  is a polynomial system whose variety is singular at  $\mathbf{x}$ . Given any system  $\mathbf{u}$ , our task is to solve the equations  $\mathbf{x} + \mathbf{c} = \mathbf{u}$  for  $(\mathbf{x}, \mathbf{c}) = (\mathbf{x}, \mathbf{u} - \mathbf{x}) \in N_X$ . The desired distance from  $\mathbf{u}$  to  $\Sigma$  is the minimum of  $\|\mathbf{x}\|$  solutions to these equations. In symbols, we have

$$\text{dist}(\mathbf{u}, \Sigma) = \min_{\mathbf{c} \in \Sigma} \|\mathbf{u} - \mathbf{c}\| = \min_{(\mathbf{x}, \mathbf{u} - \mathbf{x}) \in N_X} \|\mathbf{x}\|; \quad (9.20)$$

Here, taking the minimum happens over all real points. Of course, the minimum value (and even the ED degree) depend on the metric that we choose. In the previous section we took the Bombieri-Weyl metric.

**Example 9.25** ( $n = 1, d = 3$ ) Suppose we are given a polynomial  $\mathbf{u} = u_0 + u_1t + u_2t^2 + u_3t^3$  with three distinct complex roots. Here  $\mathcal{A} = \{0, 1, 2, 3\}$ , so  $X = X_\mathcal{A}$  is the twisted cubic curve in  $\mathbb{P}^3$ . The discriminant  $X^\vee = \Sigma$  is defined by  $\Delta = c_1^2c_2^2 - 4c_0c_2^3 - 4c_1^3c_3 + 18c_0c_1c_2c_3 - 27c_0^2c_3^2$ . We consider the following problem:

$$\text{minimize } \|\mathbf{u} - \mathbf{c}\|^2 = (u_0 - c_0)^2 + (u_1 - c_1)^2 + (u_2 - c_2)^2 + (u_3 - c_3)^2 \quad \text{subject to } \mathbf{c} \in \Sigma.$$

The conormal variety  $N_X$  is a 4-dimensional affine variety in  $\mathbb{C}^8$ . Its prime ideal equals

$$\langle x_1^2 - x_0x_2, x_1x_2 - x_0x_3, x_2^2 - x_1x_3, \Delta, \\ c_0x_0 - c_2x_2 - 2c_3x_3, c_1x_1 + 2c_2x_2 + 3c_3x_3, 3c_0x_1 + 2c_1x_2 + c_2x_3, c_1x_0 + 2c_2x_1 + 3c_3x_2 \rangle.$$

The equation  $\mathbf{x} + \mathbf{c} = \mathbf{u}$  has 7 complex solutions in  $N_X$ , since  $X$  has ED degree 7. We seek the one that minimizes  $\|\mathbf{x}\|^2$ . For a numerical example, consider the polynomial  $3t^3 - 26t^2 + 61t - 30$ , which is represented by the vector  $\mathbf{u} = (-30, 61, -26, 3)$ . We solve the equations  $\mathbf{x} + \mathbf{c} = \mathbf{u}$  and  $(\mathbf{x}, \mathbf{c}) \in N_X$ . Among

the seven complex solutions, three are real. The optimal value is  $\|\mathbf{x}\|^2 = 0.0251866$ , and this is attained by  $\mathbf{c} = \mathbf{u} - \mathbf{x} = (-29.997, 61.011, -25.959, 3.153)$ . In larger cases, where computing the ideal of  $N_X$  is infeasible, we can still solve the ED problem with the monomial parametrization of  $X$ . In our example:

$$\text{minimize } \sum_{i=0}^3 (st^i)^2 \text{ over all critical points of } (s, t) \mapsto \sum_{i=0}^3 (u_i - st^i)^2 \quad (9.21)$$

The optimal point is  $(s, t) = (-0.002824, 3.78442)$ , and the optimal value is indeed 0.0251866.  $\diamond$

We now turn to the general case  $n > 1$ . The toric variety  $X$  has dimension  $2n-1$ , and it is associated with the Cayley configuration of  $\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_n$ . To define this, we introduce new variables  $y_1, y_2, \dots, y_n$ . We encode our polynomial system  $F = (f_1, \dots, f_n)$  into the single polynomial

$$\psi(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n \sum_{\mathbf{a} \in \mathcal{A}} c_{i,\mathbf{a}} y_i \mathbf{x}^\mathbf{a}.$$

The toric variety  $X = X_{\mathcal{A}}$  is parametrized by all monomials  $y_i \mathbf{x}^\mathbf{a}$ , where  $i = 1, \dots, n$  and  $\mathbf{a} \in \mathcal{A}_i$ .

Consider the hypersurface  $\Omega := \{(\mathbf{x}, \mathbf{y}) \in (\mathbb{C}^*)^{2n} : \psi(\mathbf{x}, \mathbf{y}) = 0\}$ . The discriminant  $\Delta$  is a hypersurface in the space of coefficient vectors  $\mathbf{c} = (c_{i,\mathbf{a}})$ . It comprises all  $\mathbf{c}$  such that  $\Omega$  has a singular point. Therefore, the toric variety  $X$  is dual to the discriminant  $\Sigma$ . Points  $\mathbf{u}$  in the common ambient space of  $X$  and  $\Sigma$  are identified with polynomial systems (9.17) that have concrete numerical coefficients. The distance from such a system to the discriminant  $\Sigma$  can be computed as described in (9.20). Our discussion implies:

**Theorem 9.26** *The Euclidean distance from a polynomial system  $\mathbf{u}$  to the discriminant  $\Sigma$  equals the smallest norm  $\|\mathbf{x}\|$  among all points  $\mathbf{x}$  in the toric variety  $X$  that are critical for the distance to  $\mathbf{u}$ .*

We emphasize again that the statement of the corollary crucially depends on the norm that we choose. This was Bombieri-Weyl in Theorem 9.19. The points on the toric variety  $X$  that are critical for the distance to  $\mathbf{u}$  are best computed with the monomial parametrization of  $X$ . For  $n = 1$  this was shown in (9.21).

**Example 9.27 (Hyperdeterminant)** Fix  $n = 2$  and  $\mathcal{A}_1 = \mathcal{A}_2 = \{(0,0), (0,1), (1,0), (1,1)\}$ . The Cayley configuration is the 3-cube, and  $X$  is a 4-dimensional affine toric variety in  $\mathbb{C}^8$ , namely the cone over the Segre variety  $\mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1 \subset \mathbb{P}^7$ . Points  $\mathbf{c} \in \mathbb{C}^8$  correspond to pairs of bilinear equations in two unknowns:

$$c_{11} + c_{12}t_1 + c_{13}t_2 + c_{14}t_1t_2 = c_{21} + c_{22}t_1 + c_{23}t_2 + c_{24}t_1t_2 = 0. \quad (9.22)$$

For generic  $\mathbf{c}$ , this polynomial system has 2 solutions, and the two solutions coincide when the discriminant  $\Delta$  vanishes. By [44, Example 2.3], the discriminant for  $(\mathcal{A}_1, \mathcal{A}_2)$  is the  $2 \times 2 \times 2$  hyperdeterminant

$$\begin{aligned} \Delta = & c_{11}^2 c_{24}^2 + c_{12}^2 c_{23}^2 + c_{13}^2 c_{22}^2 + c_{14}^2 c_{21}^2 + 4c_{11}c_{14}c_{22}c_{23} + 4c_{12}c_{13}c_{21}c_{24} - 2c_{11}c_{12}c_{23}c_{24} \\ & - 2c_{11}c_{13}c_{22}c_{24} - 2c_{11}c_{14}c_{21}c_{24} - 2c_{12}c_{13}c_{22}c_{23} - 2c_{12}c_{14}c_{21}c_{23} - 2c_{13}c_{14}c_{21}c_{22}. \end{aligned}$$

Thus, given any specific polynomial system  $\mathbf{u}$  of the form (9.22), we seek its squared distance to the hyperdeterminantal hypersurface  $\Sigma = \{\Delta = 0\}$  in  $\mathbb{R}^8$ . This number is the optimal value of:

$$\begin{aligned} \text{minimize } & \sum_{i=1}^2 y_i^2 (1 + t_1^2 + t_2^2 + t_1^2 t_2^2) \text{ over all critical points of the function} \\ & (y_1, y_2, t_1, t_2) \mapsto \sum_{i=1}^2 ((u_{i1} - y_i)^2 + (u_{i2} - y_i t_1)^2 + (u_{i3} - y_i t_2)^2 + (u_{i4} - y_i t_1 t_2)^2). \end{aligned}$$

This objective function has 6 critical points over  $\mathbb{C}$ , because the ED degree is 6, by [62, Example 8.2]. The ED degree jumps to 34 if the Euclidean norm is replaced by any nearby generic quadratic form.  $\diamond$

## **Chapter 10**

## **Machine Learning**

One of the principal goals in machine learning is to learn functions that represent the relationship between data points. Suppose we are given  $r$  pairs,  $\mathcal{D} = \{(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_d, \mathbf{y}_d)\} \subset \mathbb{R}^n \times \mathbb{R}^m$ . The vectors  $\mathbf{x}_i$  are the input data, and the vectors  $\mathbf{y}_i$  are the output data. For instance, in image classification, the  $\mathbf{x}_i$  might encode images of certain objects and the  $\mathbf{y}_i$  are the respective classifier. Another popular example from generative AI is the scenario where  $\mathbf{x}_i$  and  $\mathbf{y}_i$  encode word tokens that co-occur in text data.

The goal is to find a function  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$  such that  $f(\mathbf{x}_i) \approx \mathbf{y}_i$ . The meaning of  $\approx$  rests on a *loss function*  $l(\mathbf{y}', \mathbf{y})$ . We seek to minimize the mean loss  $\ell_{\mathcal{D}}(f) := \frac{1}{d} \sum_{i=1}^d l(f(\mathbf{x}_i), \mathbf{y}_i)$  over a class of functions  $f$ . Often, the loss function comes from a metric. For instance, the *squared-error loss* is the squared Euclidean distance  $l(\mathbf{y}', \mathbf{y}) = \|\mathbf{y}' - \mathbf{y}\|^2$  in output space  $\mathbb{R}^m$ , but also other metrics are used. One usually restricts to a model, consisting of functions that are specified by  $N$  real *parameters*  $\theta = (\theta_1, \dots, \theta_N) \in \mathbb{R}^N$ . To learn a function means to compute  $\theta$  that minimizes the loss for the data  $\mathcal{D}$ .

Problems of the type described above fall under the header *supervised learning*. Here, every input data point  $\mathbf{x}_i$  has an associated output  $\mathbf{y}_i$ , also called *label*, and the goal is to find a relationship between them. By contrast, in *unsupervised learning* we are given data points  $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_d\} \subset \mathbb{R}^n$  without labels. The goal in unsupervised learning is to describe the data with as few parameters as possible. For instance, assuming the *variety hypothesis*, our goal could be to learn a variety  $X \subset \mathbb{R}^n$  that represents the data  $\mathcal{D}$ . Here, the parameters define the polynomials that cut out  $X$ . We discuss this approach in Section 10.3.

The next two sections offer a glimpse on machine learning with neural networks from a metric geometry perspective. For systematic introduction, we refer to the book *Mathematical Aspects of Deep Learning* [81].

## 10.1 Neural Networks

The model we focus on is the (*feedforward*) *neural network*. Such a network is a composition of functions

$$f := f_{L,\theta} \circ \cdots \circ f_{2,\theta} \circ f_{1,\theta},$$

where  $f_{i,\theta}$  depends on a parameter  $\theta$ . Here,  $f$  is called the *end-to-end function*. We call  $L$  the *number of layers*. The dimension of the domain of  $f_{i,\theta}$  is the *width* of the  $i$ th layer. The number of layers, their widths and the type of each layer function  $f_{i,\theta}$  are summarized as *network architecture*. The variables of the  $i$ -th layer  $f_{i,\theta}$  are called *neurons* or *nodes*. A neural network is often visualized by a graph as in Figure 10.1.

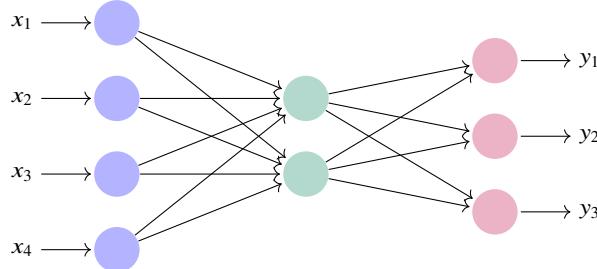


Fig. 10.1: A fully-connected neural network with two layers  $f_{1,\theta} : \mathbb{R}^4 \rightarrow \mathbb{R}^2$  and  $f_{2,\theta} : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ .

We say that the network is *fully connected* if, for every index  $i = 1, \dots, L$ , every component of the  $i$ -th layer  $f_{i,\theta}$  depends on every variable/neuron/node. Figure 10.1 shows a fully connected network. On the other hand, Figure 10.2 shows a network which is not fully connected – some of the edges are missing.

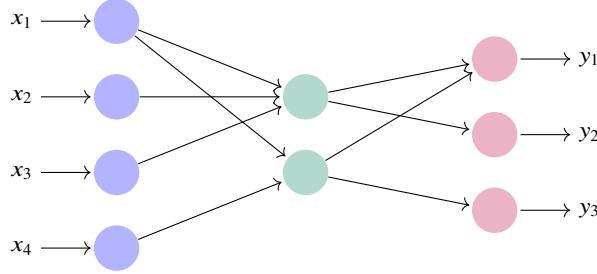


Fig. 10.2: A neural network with two layers that is not fully-connected.

Most commonly, the layer functions are compositions of an affine linear map  $\alpha_{i,\theta} : \mathbb{R}^{k_{i-1}} \rightarrow \mathbb{R}^{k_i}$  with a (typically non-linear) map  $\sigma_i : \mathbb{R} \rightarrow \mathbb{R}$  that is applied componentwise:

$$f_{i,\theta} = (\sigma_i, \dots, \sigma_i) \circ \alpha_{i,\theta}. \quad (10.1)$$

The map  $\sigma_i$  is called *activation function*. One example of an activation function is the *Rectified Linear Unit (ReLU)*, which is  $\sigma(z) = \max\{0, z\}$ . Methods from (metric or tropical) algebraic geometry can be used to study neural networks, when the activation function is algebraic (identity, polynomial, ReLU etc.).

In what follows, we consider the simplest case where the activation function is the identity. The theoretical study of neural networks is a lot more challenging if the activation function is non-linear. For instance, the end-to-end function of a ReLU neural network (i.e., all activation functions are ReLU) is piecewise linear. In fact, every piecewise linear function with finitely many pieces can be obtained from a fully-connected ReLU network [7]. ReLU end-to-end functions can be interpreted as tropical rational functions [185]. That perspective was developed in [136] to provide sharp bounds on the number of linear regions of the end-to-end functions. Another geometric perspective on ReLU networks is described in [80]. A first algebro-geometric study for networks with polynomial activation functions appears in [110].

The term *expressivity* refers to which functions neural networks with fixed architecture can express. Thus, expressivity is the study of the space  $X$  of all functions that are given by our network. This function space  $X$  is often called *neuromanifold* of the neural network architecture. Note that  $X$  is typically not a smooth manifold. The neuromanifold is the image of the network's parametrization map

$$\mu : \mathbb{R}^N \rightarrow X, \quad \theta \mapsto f_{L,\theta} \circ \dots \circ f_{2,\theta} \circ f_{1,\theta}. \quad (10.2)$$

The loss function  $\ell_D$  is defined on  $X$ . We can pull it back to the parameter space via  $\mu$ , i.e. we consider

$$\mathcal{L} : \mathbb{R}^N \xrightarrow{\mu} X \xrightarrow{\ell_D} \mathbb{R}. \quad (10.3)$$

Thus, instead of minimizing  $\ell_D$  over the neuromanifold  $X$ , training a neural network means minimizing the composition  $\mathcal{L}$  over the parameter space  $\mathbb{R}^N$ .

*Remark 10.1* The numerical uncertainty in minimizing the loss depends on condition number of  $\mathcal{L}$ . Following Chapter 9, it would be interesting to compare the condition numbers of  $\mathcal{L}$  and  $\ell_D$  and how they are related to the network's architecture. There is lots of room for metric algebraic geometry to contribute.

One of the big mysteries in machine learning theory is why training neural networks results in “nice” minima. The concrete meaning of the adjective nice varies in the literature. See also [132] for an algebro-geometric perspective on “flat” minima. Training a neural network minimizes the loss function  $\mathcal{L} = \ell_D \circ \mu$  in (10.3) on the parameter space  $\mathbb{R}^N$ . The meaningful critical points are those that come from critical

points of  $\ell_{\mathcal{D}}$  on the neuromanifold  $X$ . Formally, such *pure critical points*  $\theta$  of  $\mathcal{L}$  satisfy that  $\mu(\theta)$  is a critical point of the functional  $\ell_{\mathcal{D}}$  restricted to smooth locus of  $X$ . The network parametrization map  $\mu$  can induce additional *spurious critical points* of the function  $\mathcal{L}$ .

The following two questions play a central role in the theoretical study of learning problems:

1. How does the network architecture affect the geometry of the function space  $X$ ?
2. How does the geometry of the function space impact the training of the network?

We shall discuss these questions for two specific network architectures. In the remainder of this section, we start with the simplest class, namely networks that are linear and fully connected. Thereafter, in Section 10.2, we turn to *linear convolutional networks*. This will lead us to interesting algebraic varieties.

In a *linear fully-connected neural network*, the activation function  $\sigma_i$  in layer  $i$  is the identity and the layer functions  $f_{i,\theta}$  are arbitrary linear maps, for  $i = 1, \dots, L$ . The parameters  $\theta$  are the entries of the matrices representing the linear layer functions. The network parametrization map (10.2) specializes to

$$\mu : \mathbb{R}^{k_1 \times k_0} \times \mathbb{R}^{k_2 \times k_1} \times \cdots \times \mathbb{R}^{k_L \times k_{L-1}} \rightarrow \mathbb{R}^{k_L \times k_0}, \quad (W_1, W_2, \dots, W_L) \mapsto W_L \cdots W_2 W_1. \quad (10.4)$$

Let us first discuss expressivity of this model. The image of the parametrization map  $\mu$  consists of all matrices with rank at most  $\min(k_0, k_1, \dots, k_L)$ . Hence, the neuromanifold is the determinantal variety

$$X = \{W \in \mathbb{R}^{k_L \times k_0} \mid \text{rank}(W) \leq \min(k_0, k_1, \dots, k_L)\}.$$

If the input or output dimension is one of the minimal widths, then the neuromanifold is the whole ambient space. Equivalently, in symbols, if  $\min(k_0, k_1, \dots, k_L) = \min(k_0, k_L)$ , then  $X = \mathbb{R}^{k_L \times k_0}$ . Otherwise,  $X$  is a lower-dimensional Zariski closed subset. The singular locus of  $X$  is the variety of matrices of rank at most  $\min(k_0, k_1, \dots, k_L) - 1$ , and therefore it is the neuromanifold for a smaller network architecture.

Next, we turn to questions about optimization. The theoretical works on the optimization problem of training neural networks can be roughly grouped into *static* and *dynamic* studies. Static investigations concern the loss landscape [127] and the critical points of (10.3), while dynamic studies depend on the choice of a training algorithm, e.g., investigating its convergence. Let us first discuss static properties.

For linear fully-connected networks, the pure and spurious critical points were characterized in [35]. The critical points  $\theta$  of  $\mathcal{L}$  such that  $\mu(\theta)$  is a matrix of maximal rank are pure [35, Proposition 6]. All spurious critical points  $\theta$  correspond to lower-rank matrices  $\mu(\theta)$ . These are essentially always saddles [35, Proposition 9]. More concretely, if  $\ell_{\mathcal{D}}$  is smooth and convex, then all non-global local minima of  $\mathcal{L}$  (often called ‘‘bad’’ minima) are pure critical points. It is a common, but false, belief that linear fully-connected networks do not have bad minima. Our next result follows from [35, Proposition 10]:

**Theorem 10.2** *Consider a linear fully-connected network and a smooth and convex function  $\ell_{\mathcal{D}}$ . The function  $\mathcal{L} = \ell_{\mathcal{D}} \circ \mu$  has non-global local minima if and only if  $\ell_{\mathcal{D}}|_{\text{Reg}(X)}$  has non-global local minima.*

This result immediately implies that linear fully-connected networks *in two special settings* do not have bad minima. First, if  $X$  equals the ambient space (i.e., input or output dimension of the network is one of its minimal widths), then any convex function  $\ell_{\mathcal{D}}$  has exactly one minimum on  $X$ , and so  $\mathcal{L}$  does not have any bad minima. This is the main result from [122]. The second setting concerns the squared-error loss

$$\ell_{\mathcal{D}} : \mathbb{R}^{k_L \times k_0} \rightarrow \mathbb{R}, \quad W \mapsto \frac{1}{d} \sum_{i=1}^d \|W\mathbf{x}_i - \mathbf{y}_i\|^2, \quad (10.5)$$

for the training data  $\mathcal{D} = \{(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_d, \mathbf{y}_d)\} \subset \mathbb{R}^{k_0} \times \mathbb{R}^{k_L}$ . Writing  $X \in \mathbb{R}^{k_0 \times d}$  and  $Y \in \mathbb{R}^{k_L \times d}$  for the data matrices whose columns are  $x_i$  and  $y_j$ , the squared-error loss becomes the squared Frobenius norm

$$\ell_{\mathcal{D}}(W) = \frac{1}{d} \|WX - Y\|^2. \quad (10.6)$$

If  $XX^\top$  has full rank, then we record the data in the matrix  $U = YX^\top((XX^\top)^{\frac{1}{2}})^{-1}$ . Now, minimizing (10.6) is equivalent to minimizing the squared Euclidean distance  $\|W - U\|^2$  over all  $W \in \text{Reg}(X)$ ; see [35, Section 3.3]. By the Eckart-Young Theorem, this optimization problem has a unique local and global minimum, provided the singular values of  $U$  are distinct and positive. Hence, given many generic data pairs  $(\mathbf{x}_i, \mathbf{y}_i)$ , the squared-error loss  $\ell_{\mathcal{D}}$  has no non-global minima on  $\text{Reg}(M)$ . Theorem 10.2 now implies that the squared-error loss  $\mathcal{L}$  on the parameter space has no bad minima. This is a prominent result in machine learning, often attributed to [10] or [107]. However, the two settings where either  $X$  is a vector space or we consider the squared-error loss are rather special. Non-global local minima are expected for other loss functions and architectures where  $X$  is a proper determinantal variety; see [35, Example 13].

Now, we turn to dynamical properties. Optimization algorithms used in the training of neural networks are variations of gradient descent. After picking initial parameters  $\theta$ , gradient descent adapts the parameters successively with the goal to minimize the loss  $\mathcal{L}(\theta)$ . When training linear fully-connected networks with squared-error loss, where the data matrix  $XX^\top$  has full rank, gradient descent converges for almost all initializations (under reasonable assumptions on its step sizes) to a critical point  $\theta$  of the loss  $\mathcal{L}$  [138, Theorem 2.4]. Moreover, the matrix  $\mu(\theta) \in X$  is a global minimum of  $\ell_{\mathcal{D}}$  restricted to the smooth manifold of all matrices of the same format and rank [138, Theorem 2.6]. The authors of [138] conjecture that the matrix  $\mu(\theta)$  has maximal possible rank, so it is a smooth point on the determinantal neuromanifold  $X$ .

An essential ingredient in the convergence analysis of [138] are the *algebraic invariants* of gradient flow. The curve in parameter space traced by gradient flow is typically transcendental, but it does satisfy some algebraic relations. In other words, its Zariski closure is not the whole ambient parameter space.

**Proposition 10.3 ([8])** Consider a linear fully-connected network with parametrization (10.4). Let  $\theta(t) = (W_1(t), W_2(t), \dots, W_L(t))$  be the curve traced out by gradient flow for starting at  $t = 0$ . Then the matrices

$$W_i^\top(t) W_i(t) = W_{i-1}(t) W_{i-1}^\top(t), \quad \text{where } i = 2, 3, \dots, L,$$

remain constant for all times  $t \geq 0$ .

We discuss some practical consequences of this result. A parameter tuple  $\theta = (W_1, \dots, W_L)$  is called *balanced* if  $W_i^\top W_i = W_{i-1} W_{i-1}^\top$  for  $i \in \{2, \dots, L\}$ . Note that all matrices in a balanced tuple have the same Frobenius norm:  $\|W_1\| = \|W_2\| = \dots = \|W_L\|$ . If a linear network is initialized at a balanced tuple  $\theta(0)$ , then every tuple  $\theta(t)$  along the gradient flow curve is balanced, by Proposition 10.3. Since all matrices in  $\theta(t)$  have the same Frobenius norm, it cannot happen that one matrix converges to zero while another matrix has entries that diverge to infinity. In fact, if one matrix converges to zero, then so does  $\theta(t)$ .

Algebraic invariants of gradient flow were also studied in [113, Proposition 5.13] for linear convolutional networks (cf. Section 10.2). For ReLU and other networks see [182, Lemma 3] and [64, Theorems 2.1–2.2].

**Remark 10.4 (Nonlinear Autoencoders)** Invariants also play a crucial role in the study of attractors of autoencoders. An *autoencoder* is a composition of two neural networks: an encoder and a decoder network, such that the input dimension of the encoder equals the output dimension of the decoder. Given training data  $\mathbf{x}_1, \dots, \mathbf{x}_d$ , the composed network is typically trained by minimizing the *autoencoding loss*

$$f \mapsto \sum_{i=1}^d \|f(\mathbf{x}_i) - \mathbf{x}_i\|^2.$$

Note that  $\mathbf{x}_i$  serves both as input and as output data. It is shown in [154] that an autoencoder trained with gradient descent on a single training example  $\mathbf{x}$  (i.e.,  $d = 1$ ) memorizes that  $\mathbf{x}$  as an attractor (under suitable assumptions on the activation function and initialization). Namely,  $\mathbf{x}$  is a fixed point of the learned function  $f$  such that the sequence  $(f^i(\mathbf{y}))_{i \in \mathbb{N}}$  converges to  $\mathbf{x}$  for any  $\mathbf{y}$  in an open neighborhood of  $\mathbf{x}$ .

## 10.2 Convolutional Networks

A linear neural network parametrizes matrices that admit the structured product decomposition (10.4). In a linear convolutional neural network, the matrices  $W_i$  are generalized Toeplitz matrices. They look like

$$\begin{bmatrix} w_0 & w_1 & w_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & w_0 & w_1 & w_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & w_0 & w_1 & w_2 \end{bmatrix}. \quad (10.7)$$

Such matrices represent convolutions, and hence can be modeled by multiplication of polynomials.

A *convolution* on one-dimensional signals is a linear map depending on a *filter*  $\mathbf{w} = (w_0, \dots, w_{k-1}) \in \mathbb{R}^k$  and a *stride*  $s \in \mathbb{N}$ . It computes the inner product of the filter  $\mathbf{w}$  with parts of a given input vector  $\mathbf{x}$ , and traverses the whole vector  $\mathbf{x}$  by moving the filter  $\mathbf{w}$  through it with stride  $s$ . The formula for this is

$$\mathbb{R}^{s(m-1)+k} \rightarrow \mathbb{R}^m, \quad \mathbf{x} \mapsto \left( \sum_{j=0}^{k-1} w_j \cdot \mathbf{x}_{is+j} \right)_{i=0}^{m-1}. \quad (10.8)$$

For instance, for  $m = 3$ , the matrix (10.7) represents convolution with filter size  $k = 3$  and stride  $s = 2$ .

A *linear convolutional neural network* is the composition of  $L$  convolutions with filter sizes  $\mathbf{k} = (k_1, k_2, \dots, k_L)$  and strides  $\mathbf{s} = (s_1, s_2, \dots, s_L)$ . The activation functions in this network are all equal to the identity. The resulting end-to-end function is also a convolution with filter size  $k = k_1 + \sum_{l=2}^L (k_l - 1) \prod_{i=1}^{l-1} s_i$  and stride  $s = s_1 s_2 \cdots s_L$ . This was shown in [113, Proposition 2.2]; see also [114, Section 2]. The point is this: any product of generalized Toeplitz matrices like (10.7) is again a generalized Toeplitz matrix.

The parametrization (10.4) sends the filters of each layer to the filter of the end-to-end convolution:

$$\mu : \mathbb{R}^{k_1} \times \mathbb{R}^{k_2} \times \cdots \times \mathbb{R}^{k_L} \rightarrow \mathbb{R}^k. \quad (10.9)$$

The map  $\mu$  can be evaluated via polynomial multiplication. This is done as follows. For any positive integers  $k$  and  $t$ , we write  $\mathbb{R}[x^t]_{\leq k-1}$  for the space of univariate polynomials of degree at most  $k-1$  in the power  $x^t$ . We identify any filter of size  $k$  with the coefficient vector of such a polynomial via

$$\pi_{t,k} : \mathbb{R}^k \rightarrow \mathbb{R}[x^t]_{\leq k-1}, \quad \mathbf{w} \mapsto w_0 x^{t(k-1)} + w_1 x^{t(k-2)} + \cdots + w_{k-2} x^t + w_{k-1}.$$

The filter  $\mathbf{w}$  of the end-to-end convolution corresponds to the polynomial  $\pi_{1,k}(\mathbf{w})$  that admits a sparse factorization into polynomials corresponding to the filters  $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_L$  in the  $L$  layers:

$$\pi_{1,k}(\mu(\mathbf{w}_1, \dots, \mathbf{w}_L)) = \pi_{t_L, k_L}(\mathbf{w}_L) \cdots \pi_{t_2, k_2}(\mathbf{w}_2) \cdot \pi_{t_1, k_1}(\mathbf{w}_1), \quad \text{where } t_l := s_1 s_2 \cdots s_{l-1}. \quad (10.10)$$

The powers  $\mathbf{t} = (t_1, \dots, t_L)$  of the factors in (10.10) encode uniquely the strides  $\mathbf{s} = (s_1, \dots, s_{L-1})$  of the convolutions. The last stride  $s_L$  does not influence the filter  $\mu(\mathbf{w}_1, \dots, \mathbf{w}_L)$  of the end-to-end convolution. We summarize our discussion in the following result, which is found in [114, Proposition 2.2].

**Proposition 10.5** *The neuromanifold  $\mathcal{M}_{\mathbf{k}, \mathbf{t}}$  for a linear convolutional neural network with filter sizes  $\mathbf{k} = (k_1, k_2, \dots, k_L)$  and powers  $\mathbf{t} = (t_1, t_2, \dots, t_L)$  is parametrized by polynomial multiplication, namely*

$$\begin{aligned} \mu : \mathbb{R}[x^{t_1}]_{\leq k_1-1} \times \mathbb{R}[x^{t_2}]_{\leq k_2-1} \times \cdots \times \mathbb{R}[x^{t_L}]_{\leq k_L-1} &\rightarrow \mathbb{R}[x]_{\leq k-1}, \\ (p_1, p_2, \dots, p_L) &\mapsto p_1 p_2 \cdots p_L. \end{aligned} \quad (10.11)$$

**Example 10.6** ( $L=2, \mathbf{t} = (1, 2)$ ) We examine the cases  $\mathbf{k} = (3, 2)$  and  $\mathbf{k} = (2, 3)$ . The matrix products are

$$\begin{bmatrix} v_0 & v_1 \end{bmatrix} \begin{bmatrix} u_0 & u_1 & u_2 & 0 & 0 \\ 0 & 0 & u_0 & u_1 & u_2 \end{bmatrix} = \begin{bmatrix} u_0v_0 & u_1v_0 & u_0v_1 + u_2v_0 & u_1v_1 & u_2v_1 \end{bmatrix}.$$

$$\begin{bmatrix} v_0 & v_1 & v_2 \end{bmatrix} \begin{bmatrix} u_0 & u_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & u_0 & u_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & u_0 & u_1 \end{bmatrix} = \begin{bmatrix} u_0v_0 & u_1v_0 & u_0v_1 & u_1v_1 & u_0v_2 & u_1v_2 \end{bmatrix}.$$

These correspond to multiplying pairs of polynomials, as specified by the map  $\mu$  in Proposition 10.5:

$$(u_0x^2 + u_1x + u_2)(v_0x^2 + v_1) \quad \text{and} \quad (u_0x + u_1)(v_0x^4 + v_1x^2 + v_2). \quad (10.12)$$

The neuromanifold  $\mathcal{M}_{\mathbf{k}, \mathbf{t}}$  consists of polynomials of degree  $\leq 4$  resp. 5 that admit such a factorization.  $\diamond$

For linear convolutional networks, the powers  $\mathbf{t} = (t_1, t_2, \dots, t_L)$  satisfy that  $t_1 = 1$  and  $t_{l-1}$  divides  $t_l$  for each  $l = 2, \dots, L$ . From the algebraic perspective, it is interesting to study the polynomial multiplication map (10.11) for arbitrary positive integers  $t_l$ . Its image  $\mathcal{M}_{\mathbf{k}, \mathbf{t}}$  lives in the vector space  $\mathbb{R}[x]_{\leq k-1} \simeq \mathbb{R}^k$ , where  $k-1 = \sum_{l=1}^L t_l(k_l - 1)$ . By Tarski's Theorem on Quantifier Elimination,  $\mathcal{M}_{\mathbf{k}, \mathbf{t}}$  is a semi-algebraic set, i.e. it can be described by a Boolean combination of polynomial inequalities.

As is customary in applied algebraic geometry, we simplify our problem by replacing  $\mathcal{M}_{\mathbf{k}, \mathbf{t}}$  with its Zariski closure  $\mathcal{V}_{\mathbf{k}, \mathbf{t}}$  in the complex projective space  $\mathbb{P}^{k-1}$ . Thus, by definition,  $\mathcal{V}_{\mathbf{k}, \mathbf{t}}$  is the image of the map

$$\mathbb{P}^{k_1-1} \times \mathbb{P}^{k_2-1} \times \cdots \times \mathbb{P}^{k_L-1} \rightarrow \mathbb{P}^{k-1}, (p_1, p_2, \dots, p_L) \mapsto p. \quad (10.13)$$

**Proposition 10.7** *The map in (10.13) has no base points, so it is a morphism. Moreover, the map (10.13) is finite-to-one. If the integers  $t_1, t_2, \dots, t_L$  are pairwise distinct, then the map is generically one-to-one.*

**Proof** A product of polynomials can only be zero if one of its factors is zero. Hence, the projective multiplication map (10.13) is well-defined at all points in its domain. The map is finite-to-one because the irreducible factorization of a polynomial is unique, up to reordering the factors. If all factors have distinct powers, then the factors of a generic polynomial cannot be swapped, and the map is one-to-one.  $\square$

**Corollary 10.8** *The variety  $\mathcal{V}_{\mathbf{k}, \mathbf{t}}$  in  $\mathbb{P}^t$  has dimension  $k_1 + k_2 + \cdots + k_L - L$ , and its degree equals*

$$\frac{(k_1+k_2+\cdots+k_L-L)!}{(k_1-1)!(k_2-1)!\cdots(k_L-1)!} \quad \text{divided by the size of the general fiber of (10.13).} \quad (10.14)$$

**Proof** We consider the variety  $\mathbb{P}^{k_1-1} \times \mathbb{P}^{k_2-1} \times \cdots \times \mathbb{P}^{k_L-1}$  in its Segre embedding (see Definition 12.4). The map (10.13) is the composition of this Segre embedding followed by a linear projection. That projection has no base points by Proposition 10.7 and thus preserves dimension. Moreover, the linear projection is finite-to-one and of the same degree  $\delta$  as (10.13). Hence, the degree of its image  $\mathcal{V}_{\mathbf{k}, \mathbf{t}}$  is the degree of the Segre variety  $\mathbb{P}^{k_1-1} \times \cdots \times \mathbb{P}^{k_L-1}$  divided by  $\delta$ . The degree of the Segre variety is computed in Corollary 12.21 below. It is the multinomial coefficient in (10.14).  $\square$

**Example 10.9** If  $\mathbf{k} = (2, 2, \dots, 2)$  and  $\mathbf{t} = (1, 1, \dots, 1)$ , we simply multiply linear polynomials. Here  $k-1 = L$ . The map (10.13) is onto, by the Fundamental Theorem of Algebra, and hence  $\mathcal{V}_{\mathbf{k}, \mathbf{t}} = \mathbb{P}^t$ . The fiber has cardinality  $L!$ , since the linear factors can be reordered arbitrarily. The multinomial coefficient in (10.14) is also  $L!$ , so the degree of the image is 1. The neuromanifold  $\mathcal{M}_{\mathbf{k}, \mathbf{t}}$  is Zariski dense in  $\mathbb{R}[x]_{\leq k-1}$ . It consists of all real-rooted polynomials of degree  $\leq k-1$ , so it is an interesting semi-algebraic set.  $\diamond$

The size of the general fiber in (10.14) can be read off from  $\mathbf{k}$  and  $\mathbf{t}$ , namely it is the order of the symmetry group of the factorization (10.11). Such a symmetry arises whenever  $t_i = t_j$  for  $i \neq j$ . In this case, we can merge the factors  $\mathbb{R}[x^{t_i}]_{\leq k_i-1}$  and  $\mathbb{R}[x^{t_j}]_{\leq k_j-1}$  into a single factor  $\mathbb{R}[x^{t_i}]_{\leq k_i+k_j-2}$  without

changing the image variety  $\mathcal{V}_{\mathbf{k}, \mathbf{t}}$ . Therefore, for computing the degree or other properties of  $\mathcal{V}_{\mathbf{k}, \mathbf{t}}$ , we can always assume that  $\mathbf{t}$  has distinct coordinates.

For example, suppose that  $L = 6$ ,  $\mathbf{k} = (5, 4, 3, 7, 6, 3)$  and  $\mathbf{t} = (1, 5, 5, 5, 3, 3)$ . Then  $\mathcal{V}_{\mathbf{k}, \mathbf{t}} \subset \mathbb{P}^{80}$  has dimension  $22 = \sum_{i=1}^6 (k_i - 1)$ . After merging as described above, our new parameter vectors are  $\mathbf{k} = (5, 12, 8)$  and  $\mathbf{t} = (1, 5, 3)$ . Proposition 10.7 implies that the degree of  $\mathcal{V}_{\mathbf{k}, \mathbf{t}}$  equals  $\frac{22!}{4! 11! 7!} = 232792560$ .

Equipped with Corollary 10.8, we explore the prime ideal of  $\mathcal{V}_{\mathbf{k}, \mathbf{t}}$  in some small cases. This ideal lives in the polynomial ring  $\mathbb{Q}[c_0, c_1, \dots, c_{k-1}]$  where  $p(x) = \sum_{i=0}^{k-1} c_i x^i$  is in the codomain of (10.11). We start with Example 10.6. The two varieties are cubic threefolds, in  $\mathbb{P}^4$  resp.  $\mathbb{P}^5$ . Their prime ideals are

$$\langle c_0 c_3^2 + c_1^2 c_4 - c_1 c_2 c_3 \rangle \quad \text{and} \quad \langle c_0 c_3 - c_1 c_2, c_0 c_5 - c_1 c_4, c_2 c_5 - c_3 c_4 \rangle. \quad (10.15)$$

Thus  $\mathcal{V}_{(2,3),(1,2)} = \mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1 \subset \mathbb{P}^5$ , while  $\mathcal{V}_{(3,2),(1,2)}$  is a projection of that Segre threefold into  $\mathbb{P}^4$ .

**Example 10.10** ( $L=3, \mathbf{t} = (1, 2, 2)$ ) We extend the polynomials from (10.12) with a sparse quadratic factor:

$$\begin{aligned} p(x) &= (u_2 + u_1 x + u_0 x^2)(v_1 + v_0 x^2)(w_1 + w_0 x^2) \quad \text{for } \mathbf{k} = (3, 2, 2), \\ \text{and} \quad p(x) &= (u_1 + u_0 x)(v_2 + v_1 x^2 + v_0 x^4)(w_1 + w_0 x^2) \quad \text{for } \mathbf{k} = (2, 3, 2). \end{aligned}$$

The resulting varieties  $\mathcal{V}_{\mathbf{k}, \mathbf{t}}$  live in  $\mathbb{P}^6$  and  $\mathbb{P}^7$  respectively. We can understand them by examining the factors of  $p(x) + p(-x)$  and  $p(x) - p(-x)$ . Both varieties are 4-dimensional and they admit nice determinantal representations. The ideal of  $\mathcal{V}_{(3,2,2),(1,2,2)}$  is generated by the four  $3 \times 3$  minors of the  $3 \times 4$  matrix

$$\begin{bmatrix} c_0 & c_2 & c_4 & c_6 \\ c_1 & c_3 & c_5 & 0 \\ 0 & c_1 & c_3 & c_5 \end{bmatrix}.$$

While the parametrization for  $\mathbf{k} = (2, 3, 2)$  is not a monomial map, we find that the resulting variety  $\mathcal{V}_{(2,3,2),(1,2,2)}$  is still toric. Its prime ideal is generated by the six  $2 \times 2$  minors of the  $2 \times 4$  matrix  $\begin{bmatrix} c_0 & c_2 & c_4 & c_6 \\ c_1 & c_3 & c_5 & c_7 \end{bmatrix}$ . We conclude that  $\mathcal{V}_{(2,3,2),(1,2,2)}$  is the Segre embedding of  $\mathbb{P}^1 \times \mathbb{P}^3$  into  $\mathbb{P}^7$ .  $\diamond$

These examples suggest that our varieties  $\mathcal{V}_{\mathbf{k}, \mathbf{t}}$  are interesting objects for further study in combinatorial commutative algebra. From the perspective of metric algebraic geometry, one should pursue the concepts introduced in the previous chapters. One of these is Euclidean distance optimization. Computations show that the two varieties in Example 10.6 have ED degrees 10 and 2, while the varieties in Example 10.10 have ED degrees 23 and 2. Can we find a general formula for the ED degree of  $\mathcal{V}_{\mathbf{k}, \mathbf{t}}$  in terms of  $\mathbf{k}$  and  $\mathbf{t}$ ?

We turn our attention to describing the singular locus of the variety  $\mathcal{V}_{\mathbf{k}, \mathbf{t}}$ . As argued above, we can restrict ourselves to the case when the powers  $\mathbf{t} = (t_1, \dots, t_L)$  are pairwise distinct integers. The singular locus was determined for those varieties  $\mathcal{V}_{\mathbf{k}, \mathbf{t}}$  that come from linear convolutional networks (i.e., under the assumption  $t_{l-1} | t_l$ ) in [114, Theorem 2.8]:

**Theorem 10.11** *Let  $1 = t_1 < t_2 < \dots < t_L$  be integers such that  $t_{l-1} | t_l$  for all  $l = 2, \dots, L$ . The singular locus of  $\mathcal{V}_{\mathbf{k}, \mathbf{t}}$  consists of all its proper subvarieties  $\mathcal{V}_{\mathbf{k}', \mathbf{t}'}$ .*

This result says that the neuromanifold is singular along smaller neuromanifolds of convolutional networks with the same strides. Hence, as in the case of fully-connected linear networks, discussed in Section 10.1, the singular locus of the neuromanifold is parametrized by smaller network architectures.

In the remainder of this section, we discuss critical points of training linear convolutional networks under a loss  $\mathcal{L}$ . For that, we have to return to investigating the semi-algebraic neuromanifold  $\mathcal{M}_{\mathbf{k}, \mathbf{t}}$ . This set is closed in the Euclidean topology, but describing its Euclidean relative boundary is a challenging problem, studied in [114, Section 6].

For convolutional networks where all strides are one (i.e.,  $\mathbf{t} = (1, \dots, 1)$ ), the neuromanifold  $\mathcal{M}_{\mathbf{k}, \mathbf{t}}$  is a full-dimensional semi-algebraic subset of the ambient space  $\mathbb{R}^k$ . Here, critical points of the loss  $\mathcal{L}$

often correspond to points on the Euclidean boundary of  $\mathcal{M}_{k,t}$ . Such points are critical points of the network parametrization map  $\mu$ .

This stands in sharp contrast to convolutional networks where all strides are larger than one (i.e.,  $t_1 < t_2 < \dots$ ) and  $\mathcal{M}_{k,t}$  is a lower-dimensional subset of  $\mathbb{R}^k$ . The following results from [114, Theorem 2.11].

**Theorem 10.12** *Consider the parametrization map  $\mu$  in (10.11) with integers  $1 = t_1 < t_2 < \dots < t_L$  such that  $t_{l-1}|t_l$  for all  $l = 2, \dots, L$ . Let  $d \geq k$ , and let  $\ell_{\mathcal{D}}$  be the squared-error loss (10.5) for the end-to-end convolution  $W$ . For almost all  $d$ -tuples  $\mathcal{D}$  of training data, every critical point of  $\mathcal{L} = \ell_{\mathcal{D}} \circ \mu$  satisfies:*

1. either  $\mu(\theta) = 0$ , or
2.  $\theta$  is a regular point of  $\mu$  and  $\mu(\theta)$  lies in the smooth locus of  $\mathcal{V}_{k,t}$  and the Euclidean interior of  $\mathcal{M}_{k,t}$ .

We summarize our discussion on the training of linear networks with the squared-error loss. For fully-connected networks as in Section 10.1, the neuromanifold  $\mathcal{M}$  is a classical determinantal variety. Its Euclidean relative boundary is empty. Nevertheless, spurious critical points commonly appear; they correspond to singular points of  $\mathcal{M}$ . For convolutional networks of stride one,  $\mathcal{M}$  is semi-algebraic, Euclidean closed and full-dimensional. Here, the singular locus (of its Zariski closure) is empty, but often critical points are on its Euclidean boundary and are thus critical points of  $\mu$ . Finally, for convolutional networks with all strides larger than one,  $\mathcal{M}$  is a semi-algebraic, Euclidean closed, lower-dimensional subset. Its Euclidean relative boundary and its singular locus are usually non-trivial. Nevertheless, these loci are not relevant for training the network when using a sufficient amount of generic data: All critical points – except when a filter in one of the layers is zero – are pure and correspond to interior smooth points of  $\mathcal{M}$ .

*Remark 10.13* In this section, we studied convolutions on one-dimensional signals. Many practical neural networks, notably in image processing, use convolutions on two-dimensional signals. Higher-dimensional convolutions move a *filter tensor* through an *input tensor* of the same dimension. The composition of such convolutions corresponds to the multiplication of multivariate polynomials [113, Section 4.3].

## 10.3 Learning Algebraic Varieties

Suppose that we are given  $d$  data points  $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_d\} \subset \mathbb{R}^n$ . We assume the *variety hypothesis*. This means we assume that the points in  $\mathcal{D}$  are (possibly noisy) samples from a real algebraic variety  $X \subset \mathbb{R}^n$ . The goal is to learn the variety  $X$  from the data  $\mathcal{D}$ . The unknown variety  $X$  is represented by the ideal  $I(X)$  of all polynomials that vanish on  $X$ . Assuming  $X$  to be irreducible, we know that  $I(X)$  is prime. Otherwise  $I(X)$  is unknown. All we are given is the finite set  $\mathcal{D}$ . In this section we discuss strategies to learn an ideal  $I_{\mathcal{D}}$  which is meant to approximate  $I(X)$ . Our discussion is an invitation to the article [31].

*Remark 10.14* In some situations, one is interested in specific features of  $X$  like, for instance, its dimension or homology. In this case, learning the ideal means asking for more information than what is actually needed. Other strategies can be more efficient: Theorem 15.2 in Chapter 15 gives conditions under which one can recover the homology of  $X$  from the finite sample  $\mathcal{D}$  by computing the Čech-complex of a union of balls. For estimating the dimension of  $X$ , see [31, Section 3] and the vast literature on data dimensionality.

Thus, we now focus on learning the approximate ideal  $I_{\mathcal{D}}$ . First, we need to find an appropriate loss function. One immediate option is to use mean squared loss  $\ell_{\mathcal{D}}(f) = \frac{1}{d} \sum_{i=1}^d f(\mathbf{x}_i)^2$ . In this case, we must define a threshold  $\varepsilon \geq 0$  so that we keep only those polynomials for which  $\ell_{\mathcal{D}}(f) \leq \varepsilon$ . Moreover, since  $f \in I(X)$  if and only if  $\lambda \cdot f \in I(X)$  for every  $\lambda \in \mathbb{R}$ , we must work with some kind of normalization for  $f$ . The next example illustrates one approach to computing the approximate ideal  $I_{\mathcal{D}}$  using linear algebra.

**Example 10.15** Suppose  $n = 2$  and that we have  $d = 3$  data points  $\mathcal{D} = \{\mathbf{x}, \mathbf{y}, \mathbf{z}\} \subset \mathbb{R}^2$ . Given  $\varepsilon > 0$  we want to compute polynomials  $f \in \mathbb{R}[x_1, x_2]$  of degree at most 2 that

1. have mean squared loss at most  $\varepsilon \geq 0$ , and
2. satisfy  $f_1^2 + \dots + f_6^2 = 1$ , where  $f = f_1 x_1^2 + f_2 x_1 x_2 + f_3 x_2^2 + f_4 x_1 + f_5 x_2 + f_6$ .

For this we set up the bivariate Vandermonde matrix of degree at most 2 for our data points:

$$M_{\mathcal{D}} := \begin{bmatrix} x_1^2 & x_1 x_2 & x_2^2 & x_1 & x_2 & 1 \\ y_1^2 & y_1 y_2 & y_2^2 & y_1 & y_2 & 1 \\ z_1^2 & z_1 z_2 & z_2^2 & z_1 & z_2 & 1 \end{bmatrix} \in \mathbb{R}^{3 \times 6}. \quad (10.16)$$

The mean squared loss equals  $\frac{1}{d} \sum_{i=1}^d f(\mathbf{x}_i)^2 = \frac{1}{d} \|M_{\mathcal{D}} f\|^2$ . Suppose that  $f \in \mathbb{R}^6$  is a right-singular vector of  $M_{\mathcal{D}}$  with singular value  $\sigma$ . Then,  $\|M_{\mathcal{D}} f\|^2 = \sigma^2$ . This means one can compute the singular value decomposition of  $M_{\mathcal{D}}$  and keep those the right singular vectors of  $M_{\mathcal{D}}$  whose squared singular values are at most  $d \cdot \varepsilon$ . (Careful: this does *not* yield all polynomials with mean squared loss at most  $\varepsilon$ ). The polynomials corresponding to these singular vectors generate an ideal, which is a candidate for  $I_{\mathcal{D}}$ .  $\diamond$

The approach from Example 10.15 can be generalized to any ambient dimension  $n$  and any number of points  $d$ . The matrix in (10.16) is called a *multivariate Vandermonde matrix*. In the general setting it is a  $d \times m$ -matrix, where  $m$  is the number of monomials we are considering for our polynomials in  $I_{\mathcal{D}}$ . In Example 10.15 all monomials of degree at most 2 are used. In general, the choice of monomials is a hyperparameter for this approach. This means that we have to specify the monomials before solving for  $f$ .

In Example 10.15 we used the singular value decomposition for computing polynomials that generate our ideal  $I_{\mathcal{D}}$ . Two alternative methods are discussed in [31, Section 5], namely the QR-decomposition and row-echelon, also for the Vandermonde matrix. All three methods have their advantages and disadvantages.

In other situations one might have prior knowledge on the polynomials that generate  $I(X)$ . This might come from geometric considerations (as in [31, Section 2]). These might suggest determinantal representations, or torus actions on  $X$ , which ensure that  $I(X)$  is homogeneous with respect to some multigrading. This amounts to information about sparsity. Sometimes, we have a priori bounds for the degrees of the generators of  $I(X)$ . The “variety of varieties” in [31, Section 2.2] suggests many options.

In order to search for sparse polynomials, we can utilize methods from *compressed sensing*. This field is concerned with computing sparse solutions to underdetermined systems of linear equations. Moreover, polynomials coming from specific applications often have rational or even integer coefficients, so one might seek polynomials that also have this property. For this, finite field techniques can be very useful.

For the normalization of the polynomials in Example 10.15, we used the Euclidean norm of the vector of coefficients. An alternative approach is to fix a monomial ordering on  $\mathbb{R}[x_1, \dots, x_n]$  (see Chapter 3), and then requiring the leading coefficient of every  $f$  to be 1. This leads to the following definition.

**Definition 10.16** Let  $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_d\} \subset \mathbb{R}^n$  be a finite sample of points and  $\varepsilon \geq 0$ . The  $\varepsilon$ -approximate vanishing ideal over  $\mathcal{D}$  is  $I_{\mathcal{D}}^{\varepsilon} := \langle \{f \mid \frac{1}{d} \sum_{i=1}^d f(\mathbf{x}_i)^2 \leq \varepsilon \text{ and the leading coefficient of } f \text{ is 1}\} \rangle$ .

We always have  $I(X) \subseteq I_{\mathcal{D}}^{\varepsilon}$ . In the limit, when the sample  $\mathcal{D}$  consists of all points in  $X$  and  $\varepsilon = 0$ , we have  $I_X^0 = I(X)$ . However, it is not known under which conditions the dimension of the variety of  $I_{\mathcal{D}}^{\varepsilon}$  is equal to the dimension of  $X$ , or how their degrees are related (if at all). There is a vast literature on the approximate vanishing ideal and how to compute it. We refer to [152] and the references therein.

Wirth and Pokutta [152] present a novel algorithm for computing the approximate vanishing ideal  $I_{\mathcal{D}}^{\varepsilon}$ , given any value for the tolerance  $\varepsilon \geq 0$ . By contrast to the approach from Example 10.15, for their algorithm they do not have to specify monomials; i.e., monomials are not hyperparameters. The key idea is to use *border bases* of ideals [108], which are generalizations of Gröbner bases that have more desirable numerical properties. Moreover, the algorithm of [152] returns sparse generators for the ideal  $I_{\mathcal{D}}^{\varepsilon}$ .

Both approaches presented above use the mean squared loss that comes from evaluating polynomials. A more geometric choice for the loss would be  $\ell_{\mathcal{D}} = \frac{1}{d} \sum_{i=1}^d d(\mathbf{x}_i, X)^2$ , where  $d(\mathbf{x}, X)$  is the Euclidean distance from  $\mathbf{x}$  to  $X$ . This idea has not been worked out, neither from a theoretical nor from an algorithmic perspective. This desirable aim underscores the relevance of the ED problem (2.1) for machine learning.

*Remark 10.17* In Chapter 15 we present methods for sampling points from algebraic varieties given their equations. Thus, once we have learned equations, we can set up a *generative model* that produces synthetic data. For instance, Figure 15.4 shows points that have been generated near the Trott curve. For enlightening varieties in higher dimensions see [31, Section 2.2]. One data set  $\mathcal{D}$  to start with is [31, Example 2.13].

We now shift gears. In the remainder of this section, we discuss how machine learning can be helpful for theoretical research in algebraic geometry. Several articles have explored this theme, and we shall offer a brief guide to this emerging literature. However, it is not clear yet what type of information can be learned and which type of tasks can be solved by training neural networks. Despite its massive success for classification tasks (e.g., “is there a cat on this picture?”), machine learning with neural networks has not improved explicitly geometric tasks such as solving systems of polynomial equations. For instance, the problem of reconstructing 3D scenes from images taken by unknown cameras (see Section 13.3) is essentially equivalent to solving certain polynomial equation systems. There have been numerous attempts to solve that problem with machine learning methods. However, none of them have been as successful as traditional techniques involving symbolic computations with Gröbner bases or resultants [163, 186].

The works that have used machine learning techniques to answer questions in algebraic geometry come roughly in two flavors. On the one hand, machine learning has been used to directly compute geometric properties, e.g. of Hilbert series [11], of irreducible representations [49], or numerical Calabi-Yau (Ricci flat Kähler) metrics [61]. Those approaches trade off the reliability of the output with performance, and this can yield valuable insights into problem instances that lie outside of the scope of traditional techniques.

On the other hand, many algebro-geometric algorithms depend on a heuristic that has to be chosen by the user and that might heavily influence their performance. For instance, to compute a Gröbner basis, a monomial ordering has to be chosen. Machine learning has been successful at predicting such a heuristic, which is then used to speeds up the computation using traditional algorithms. In that way the performance can be enhanced without compromising the reliability of the final output. This approach has been used to speed up Buchberger’s Algorithm, by learning S-pair selection strategies, [146] and Cylindrical Algebraic Decomposition by learning a variable ordering and by exploring Gröbner basis preconditioning [102].

In a similar spirit, the computation of periods of hypersurfaces is enhanced in [89]. The authors study pencils of hypersurfaces, and they use neural networks to predict the complexity of the Gauss-Manin connection for the period matrix along the pencil. Based on that prediction, they explore all smooth quartic surface in  $\mathbb{P}^3$  whose polynomial is a sum of five monomials, and they guess when their periods are computable by Gauss-Manin. This leads them to determine the periods of 96% of those surfaces.

Although neural networks have not shown a great potential for solving systems of polynomials, they can be trained to predict their number of real solutions [18, 32]. Such a prediction might be used by homotopy methods that track real solutions only. This assumes one has pre-computed a starting system for each possible number of real solutions, and – ideally – one for each chamber of the real discriminant.

That would yield a reliable numerical computation that requires less computation time since it tracks fewer paths. A more drastic approach has been taken in [100] where the authors learn a *single* starting solution for a real homotopy that has good chances to reach a good solution of the desired target system. That way of computing produces a less reliable solution, but since they propose their method as part of a random sample consensus (RANSAC) scheme, bad solutions can be detected and disregarded. Since tracking a single solution can be very fast, they can simply repeat their approach for each bad solution.



## **Chapter 11**

### **Maximum Likelihood**

This book started out with the problem of minimizing the Euclidean distance from a data point  $\mathbf{u}$  to a model  $X$  that is given by polynomial equations. We studied the analogous problem in the setting of algebraic statistics [173], where the model  $X$  represents a family of probability distributions, and we used the Wasserstein metric to measure the distances from  $\mathbf{u}$  to  $X$ . In this chapter we stay with statistical models but we now use Kullback-Leibler divergence and likelihood inference instead of Wasserstein distance.

## 11.1 Kullback-Leibler Divergence

The two scenarios of most interest for statisticians are Gaussian models and discrete models. We start with discrete models, where we take the state space is the finite set  $\{0, 1, \dots, n\}$ . The simplex of all probability distributions on this state space equals

$$\Delta_n = \{\mathbf{p} = (p_0, p_1, \dots, p_n) \in \mathbb{R}^{n+1} : p_0 + p_1 + \dots + p_n = 1 \text{ and } p_0, p_1, \dots, p_n > 0\}. \quad (11.1)$$

Given two probability distributions  $\mathbf{q}$  and  $\mathbf{p}$  in  $\Delta_n$ , their *Kullback-Leibler (KL) divergence* is defined as

$$D_{\text{KL}}(\mathbf{q} \parallel \mathbf{p}) = \sum_{i=0}^n q_i \cdot \log(q_i/p_i). \quad (11.2)$$

This function is not symmetric in its two arguments, i.e. we have  $D_{\text{KL}}(\mathbf{q} \parallel \mathbf{p}) \neq D_{\text{KL}}(\mathbf{p} \parallel \mathbf{q})$  in general. Nevertheless, we interpret KL divergence as a kind of metric on the open simplex  $\Delta_n$ .

**Lemma 11.1** *The KL divergence is nonnegative and it is zero if and only if the two distributions agree. In symbols,  $D_{\text{KL}}(\mathbf{q} \parallel \mathbf{p}) \geq 0$  for all  $\mathbf{p}, \mathbf{q} \in \Delta_n$ , and equality holds if and only if  $\mathbf{p} = \mathbf{q}$ .*

**Proof** We use the calculus fact that the function  $x \mapsto (x - 1) - \log(x)$  is nonnegative for  $x \in \mathbb{R}_{>0}$  and its only zero occurs at  $x = 1$ . Hence the sum in (11.2) is bounded below as follows:

$$D_{\text{KL}}(\mathbf{q} \parallel \mathbf{p}) = - \sum_{i=0}^n q_i \cdot \log(p_i/q_i) \geq - \sum_{i=0}^n q_i \cdot (p_i/q_i - 1) = \sum_{i=0}^n p_i - \sum_{i=0}^n q_i = 1 - 1 = 0.$$

Moreover, equality holds if and only if  $p_i/q_i = 1$  for all indices  $i$ .  $\square$

Our model is a subset  $X$  of  $\Delta_n$  defined by homogeneous polynomial equations. As before, for venturing beyond linear algebra, we identify  $X$  with its Zariski closure in complex projective space  $\mathbb{P}^n$ .

We shall present the algebraic approach to maximum likelihood estimation (MLE). Our sources include [43, 65, 94, 105, 106, 173] and references therein. Suppose we are given  $N$  i.i.d. samples. These are summarized in the *data vector*  $\mathbf{u} = (u_0, u_1, \dots, u_n)$  where  $u_i$  is the number of samples that were in state  $i$ . Thus the sample size is  $N = u_0 + u_1 + \dots + u_n$ . The associated log-likelihood function equals

$$\ell_{\mathbf{u}} : \Delta_n \rightarrow \mathbb{R}, \mathbf{p} \mapsto u_0 \cdot \log(p_0) + u_1 \cdot \log(p_1) + \dots + u_n \cdot \log(p_n).$$

Performing MLE for the model  $X$  means solving the following optimization problem:

$$\text{Maximize } \ell_{\mathbf{u}}(\mathbf{p}) \text{ subject to } \mathbf{p} \in X. \quad (11.3)$$

Viewed through the lens of metric algebraic geometry, this problem amounts to minimizing a certain distance, namely KL divergence, to the variety  $X$ . Namely, given a data vector  $\mathbf{u}$  with  $u_i > 0$  for all  $i$ , we write  $\mathbf{q} = \frac{1}{N}\mathbf{u}$  for the corresponding empirical distribution in  $\Delta_n$ .

*Remark 11.2* The maximum likelihood estimation problem (11.3) is equivalent to:

$$\text{Minimize } D_{\text{KL}}(\mathbf{q} \parallel \mathbf{p}) \text{ subject to } \mathbf{p} \in X. \quad (11.4)$$

This holds because the KL divergence can be rewritten as the entropy of the empirical distribution  $\mathbf{q}$  minus the log-likelihood function:  $D_{\text{KL}}(\mathbf{q} \parallel \mathbf{p}) = \sum_{i=0}^n q_i \log(q_i) - \frac{1}{N} \ell_{\mathbf{u}}(\mathbf{p})$ .

As before, we identify the model  $X$  with a projective variety in  $\mathbb{P}^n$ . The objective function in the optimization problem (11.3) involves logarithms and it is not an algebraic function. However, each of its partial derivatives is a rational function, and therefore we can study this problem using algebraic geometry.

## 11.2 Maximum Likelihood Degree

We fix a real projective variety  $X$  in  $\mathbb{P}^n$ , and we consider the the optimization problem in (11.3) or (11.4). The *maximum likelihood degree (ML degree)* of  $X$  is defined to be the number of complex critical points of this optimization problem for generic data  $\mathbf{u}$ . For arbitrary fixed data  $\mathbf{u}$ , the optimal solution is denoted  $\hat{\mathbf{p}}$  and called the *maximum likelihood estimate* of the model  $X$  for the data  $\mathbf{u}$ . Thus ML degree is the analogue to ED degree, when now KL divergence replaces Euclidean distance.

The critical equations for (11.3) are similar to those of ED problem. Let  $I_X = \langle f_1, \dots, f_k \rangle$  be the homogeneous ideal of the model  $X$ . In addition, we consider the inhomogeneous linear polynomial  $f_0 := p_0 + p_1 + \dots + p_n - 1$ . Let  $\mathcal{J} = (\partial f_i / \partial p_j)$  denote the Jacobian matrix of size  $(k+1) \times (n+1)$  for these polynomials, and set  $c = \text{codim}(X)$ . The *augmented Jacobian*  $\mathcal{AJ}$  is obtained from  $\mathcal{J}$  by prepending one more row, namely the gradient of the objective function

$$\nabla \ell_{\mathbf{u}} = (u_0/p_0, u_1/p_1, \dots, u_n/p_n).$$

To obtain the critical equations, enlarge  $I_X$  by the  $(c+2) \times (c+2)$  minors of the  $(k+2) \times (n+1)$  matrix  $\mathcal{AJ}$ , then clear denominators, and remove extraneous components by saturation.

**Example 11.3 (Space curves)** Let  $n = 3$  and  $X$  the curve in  $\Delta_3$  defined by two general polynomials  $f_1$  and  $f_2$  of degrees  $d_1$  and  $d_2$  in  $p_0, p_1, p_2, p_3$ . The augmented Jacobian matrix is

$$\mathcal{AJ} = \begin{pmatrix} u_0/p_0 & u_1/p_1 & u_2/p_2 & u_3/p_3 \\ 1 & 1 & 1 & 1 \\ \partial f_1 / \partial p_0 & \partial f_1 / \partial p_1 & \partial f_1 / \partial p_2 & \partial f_1 / \partial p_3 \\ \partial f_2 / \partial p_0 & \partial f_2 / \partial p_1 & \partial f_2 / \partial p_2 & \partial f_2 / \partial p_3 \end{pmatrix}. \quad (11.5)$$

Clearing denominators amounts to multiplying the  $i$ th column by  $p_i$ , so the determinant contributes a polynomial of degree  $d_1 + d_2 + 1$  to the critical equations. Since the codimension of  $X$  equals  $c = 2$ , we need to take the  $4 \times 4$  minors of  $\mathcal{AJ}$ . The generators of  $I_X$  have degrees  $d_1$  and  $d_2$  respectively. We therefore conclude that the ML degree of  $X$  equals  $d_1 d_2 (d_1 + d_2 + 1)$ .

The following general upper bound on the ML degree is established in [94, Theorem 5].

**Proposition 11.4** Let  $X$  be a model of codimension  $c$  in the probability simplex  $\Delta_n$  whose ideal  $I_X$  is generated by polynomials  $f_1, f_2, \dots, f_c, \dots, f_k$  of degrees  $d_1 \geq d_2 \geq \dots \geq d_c \geq \dots \geq d_k$ . Then

$$\text{MLdegree}(X) \leq d_1 d_2 \cdots d_c \cdot \sum_{i_1+i_2+\dots+i_c \leq n-c} d_1^{i_1} d_2^{i_2} \cdots d_c^{i_c}. \quad (11.6)$$

Equality holds when  $X$  is a generic complete intersection of codimension  $c$  (hence  $c = k$ ).

We next present a more precise formula. For the ED degree, the polar degrees in  $\mathbb{P}^n$  were used to give such a formula. For the ML degree, we shall use the Euler characteristic instead.

Given our variety  $X$  in the complex projective space  $\mathbb{P}^n$ , and let  $X^\circ$  be the open subset of  $X$  that is obtained by removing the hyperplane arrangement  $\{p_0 p_1 \cdots p_n (\sum_{i=0}^n p_i) = 0\}$ . We recall from [104, 105] that a *very affine variety* is a closed subvariety of an algebraic torus  $(\mathbb{C}^*)^r$ . In our setting, the open set  $X^\circ$  is a very affine variety, with  $r = n + 2$ . The following formula works for any very affine variety.

**Theorem 11.5** Suppose that the very affine variety  $X^\circ$  is non-singular. The ML degree of the model  $X$  equals the signed Euler characteristic  $(-1)^{\dim(X)} \cdot \chi(X^\circ)$  of the manifold  $X^\circ$ .

**Proof (and Discussion)** This was proved under additional assumptions in [43, Theorem 19], and in full generality in [104, Theorem 1]. If the very affine variety  $X^\circ$  is singular, then the Euler characteristic can be replaced by the Chern-Schwartz-MacPherson class, as shown in [104, Theorem'2].  $\square$

The optimal solution to our optimization problem (11.3)-(11.4) in the statistical model  $X^\circ \cap \Delta_n = X \cap \Delta_n$  is denoted  $\hat{\mathbf{p}}$ . This point is called the *maximum likelihood estimate (MLE)* for the given data  $\mathbf{u}$  and the model  $X$ . The ML degree measures the algebraic complexity of the MLE. Theorem 11.5 says that the ML degree is a topological invariant. Varieties  $X$  for which the ML degree is equal to one are of special interest, both statistically and geometrically. For a model  $X$  to have ML degree one means that the MLE  $\hat{\mathbf{p}}$  is a rational function of the data  $\mathbf{u}$ . Here are two natural examples where this happens.

**Example 11.6** ( $n = 3$ ) The independence model for two binary random variables is a quadratic surface  $X$  in the tetrahedron  $\Delta_3$ . This model is described by the constraints

$$\det \begin{bmatrix} p_0 & p_1 \\ p_2 & p_3 \end{bmatrix} = 0 \quad \text{and} \quad p_0 + p_1 + p_2 + p_3 = 1 \quad \text{and} \quad p_0, p_1, p_2, p_3 > 0.$$

Consider data  $\mathbf{u} = \begin{bmatrix} u_0 & u_1 \\ u_2 & u_3 \end{bmatrix}$  of sample size  $|u| = u_0 + u_1 + u_2 + u_3$ . The ML degree of the surface  $X$  equals one because the MLE  $\hat{\mathbf{p}}$  is a rational function of the data. To be precise, the coordinates of the MLE  $\hat{\mathbf{p}}$  are

$$\begin{aligned} \hat{p}_0 &= |u|^{-2}(u_0+u_1)(u_0+u_2), & \hat{p}_1 &= |u|^{-2}(u_0+u_1)(u_1+u_3), \\ \hat{p}_2 &= |u|^{-2}(u_2+u_3)(u_0+u_2), & \hat{p}_3 &= |u|^{-2}(u_2+u_3)(u_1+u_3). \end{aligned} \tag{11.7}$$

In words, we multiply the row sums with the column sums in the empirical distribution  $\frac{1}{|u|}\mathbf{u}$ .

Here is another simple model where the MLE is a rational function in the data.

**Example 11.7** ( $n = 2$ ) Given a biased coin, we perform the following experiment: *Flip a biased coin. If it shows heads, flip it again.* The outcome of this experiment is the number of heads: 0, 1 or 2.

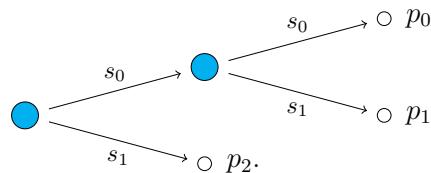


Fig. 11.1: Probability tree that describes the coin toss model in Example 11.7.

If  $s$  denotes the bias of our coin, then the model is the parametric curve  $X$  given by

$$(0, 1) \rightarrow X \subset \Delta_2, \quad s \mapsto (s^2, s(1-s), 1-s).$$

This model is the conic  $X = V(p_0 p_2 - (p_0 + p_1) p_1) \subset \mathbb{P}^2$ . Its MLE is given by the formula

$$(\hat{p}_0, \hat{p}_1, \hat{p}_2) = \left( \frac{(2u_0 + u_1)^2}{(2u_0 + 2u_1 + u_2)^2}, \frac{(2u_0 + u_1)(u_1 + u_2)}{(2u_0 + 2u_1 + u_2)^2}, \frac{u_1 + u_2}{2u_0 + 2u_1 + u_2} \right). \quad (11.8)$$

Since the coordinates of  $\hat{p}$  are rational functions, the ML degree of  $X$  is equal to one.

The following theorem explains what we saw in equations (11.7) and (11.8):

**Theorem 11.8** *If  $X \subset \Delta_n$  is a model of ML degree one, so that  $\hat{\mathbf{p}}$  is a rational function of  $\mathbf{u}$ , then each coordinate  $\hat{p}_i$  is an alternating product of linear forms with positive coefficients.*

**Proof (and Discussion)** This was proved in the setting of arbitrary complex very affine varieties by Huh in [105]. It was adapted to real algebraic geometry and hence to statistical models in [65]. These two articles offer precise statements via Horn uniformization for  $A$ -discriminants [76], i.e. hypersurfaces dual to toric varieties. For additional information we refer to [106, Corollary 3.12].  $\square$

Models given by rank constraints on matrices and tensors are particularly important in applications, since these represent conditional independence. Consider two random variables, having  $n_1$  and  $n_2$  states respectively, which are conditionally independent, given a hidden random variable with  $r$  states. In algebraic geometry, this model is the variety  $X_r$  in  $\mathbb{P}^{n_1 n_2 - 1}$  that is defined by the  $(r+1) \times (r+1)$  minors of an  $n_1 \times n_2$  matrix  $(p_{ij})$ . The ML degree of this rank  $r$  model was first studied by Hauenstein, Rodriguez and Sturmfels in [88], who obtained the following results using methods from numerical algebraic geometry.

**Proposition 11.9** *For small values of  $n_1$  and  $n_2$ , the ML degrees of low rank models  $X_r$  are*

|         | $(n_1, n_2) = (3, 3)$ | $(3, 4)$ | $(3, 5)$ | $(4, 4)$ | $(4, 5)$ | $(4, 6)$ | $(5, 5)$ |
|---------|-----------------------|----------|----------|----------|----------|----------|----------|
| $r = 1$ | 1                     | 1        | 1        | 1        | 1        | 1        | 1        |
| $r = 2$ | 10                    | 26       | 58       | 191      | 843      | 3119     | 6776     |
| $r = 3$ | 1                     | 1        | 1        | 191      | 843      | 3119     | 61326    |
| $r = 4$ |                       |          |          | 1        | 1        | 1        | 6776     |
| $r = 5$ |                       |          |          |          |          |          | 1        |

Every entry in the  $r = 1$  row is 1 because the MLE for the independence model is a rational function in the data  $(u_{ij})$ . One finds  $\hat{\mathbf{p}} = (\hat{p}_{ij})$  by multiplying the column vector of row sums of  $\mathbf{u}$  with the row vector of column sums of  $\mathbf{u}$ , and then dividing by  $|\mathbf{u}|^2$ , as shown in (11.7). The other entries are more interesting, and they give precise information on the algebraic complexity of minimizing the Kullback-Leibler distance from a given data matrix  $\mathbf{u}$  to the conditional independence model  $X_r$ . Here is an example taken from [88].

**Example 11.10** ( $n_1 = n_2 = 5$ ) Following [88, Example 7], we consider the data

$$\mathbf{u} = \begin{pmatrix} 2864 & 6 & 6 & 3 & 3 \\ 2 & 7577 & 2 & 2 & 5 \\ 4 & 1 & 7543 & 2 & 4 \\ 5 & 1 & 2 & 3809 & 4 \\ 6 & 2 & 6 & 3 & 5685 \end{pmatrix}.$$

For  $r = 2$  and  $r = 4$ , this instance of our MLE problem has the expected number of 6776 distinct complex critical points. In both cases, 1774 of these are real and 90 of these are real and positive. This illustrates the last statement in Theorem 11.11 below. The number of local maxima for  $r = 2$  equals 15, and the number of local maxima for  $r = 4$  equals 6. For  $r = 3$ , we have 61326 critical points, of which 15450 are real. Of these, 362 are positive and 25 are local maxima. We invite our readers to critically check these claims, by running software for solving polynomial equations, such as `HomotopyContinuation.jl`.

The columns of the table in (11.9) exhibit an obvious symmetry. This was conjectured in [88], and it was proved by Draisma and Rodriguez in their article [63] on maximum likelihood duality. We now state their result. Given an  $n_1 \times n_2$  matrix  $\mathbf{u}$ , we write  $\Omega_{\mathbf{u}}$  for the matrix whose  $(i, j)$  entry equals

$$\frac{u_{ij}u_{i+j}}{(u_{++})^3}.$$

In the following theorem, the symbol  $\star$  denotes the Hadamard product (or entrywise product) of two matrices. All matrices  $\mathbf{p}_i$  and  $\mathbf{q}_i$  have format  $n_1 \times n_2$  and they have complex entries.

**Theorem 11.11** *Fix  $n_1 \leq n_2$  and  $\mathbf{u}$  an  $n_1 \times n_2$ -matrix with strictly positive integer entries. There exists a bijection between the complex critical points  $\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_s$  of the likelihood function for  $\mathbf{u}$  on  $X_r$  and the complex critical points  $\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_s$  on  $X_{n_1-r+1}$  such that*

$$\mathbf{p}_1 \star \mathbf{q}_1 = \mathbf{p}_2 \star \mathbf{q}_2 = \cdots = \mathbf{p}_s \star \mathbf{q}_s = \Omega_{\mathbf{u}}. \quad (11.10)$$

*In particular, this bijection preserves reality, positivity, and rationality of the critical points.*

This result represents a multiplicative version of the duality we encountered in our study of ED degrees. Recall that the ED degree of any projective variety  $X$ , when viewed as a cone in affine space, equals that of its dual variety  $X^\vee$ . Under some genericity assumption, this common ED degree is the sum of the polar degrees, which arises from the conormal variety  $N_X = N_{X^\vee}$ . By “multiplicative” we mean that  $\mathbf{u}_i/\mathbf{p}_i$  instead of  $\mathbf{u}_i - \mathbf{p}_i$  appears in the first row of the augmented Jacobian matrix.

It is a challenge in intersection theory and singularity theory to find general formulas for the ML degrees in Proposition 11.9. This problem was solved for  $r = 2$  by Rodriguez and Wang in [159]. They give a recursive formula in [159, Theorem 4.1], and they present impressive values in [159, Table 1]. They unravel the recursion, and they obtain the explicit formulas for the ML degree of conditional independence in many cases. In particular, they obtain the following result which had been stated as a conjecture in [88].

**Theorem 11.12 (Rodriguez-Wang [159])** *Consider the variety  $X_2 \subset \mathbb{P}^{3n-1}$  whose points are the  $3 \times n$  matrices of rank  $\leq 2$ . The ML degree of this variety equals  $2^{n+1} - 6$ .*

### 11.3 Scattering Equations

We now turn to a connection between algebraic statistics and particle physics that was developed in [171]. The context is scattering amplitudes, where the critical equations for (11.3)-(11.4) are known as *scattering equations*. We consider the *CEGM model*, due to Cachazo and his collaborators [39, 40]. The role of the data vector  $\mathbf{u}$  is played in physics by the *Mandelstam invariants*. This theory rests on the space  $X^o$  of  $m$  labeled points in general position in  $\mathbb{P}^{k-1}$ , up to projective transformations. Consider the Grassmannian  $\text{Gr}(k, m)$  in its Plücker embedding into  $\mathbb{P}^{\binom{m}{k}-1}$ . The torus  $(\mathbb{C}^*)^m$  acts on  $\text{Gr}(k, m)$  by scaling the columns of  $k \times m$  matrices representing subspaces. Let  $\text{Gr}(k, m)^o$  be the open Grassmannian where all Plücker coordinates are nonzero. The CEGM model is the  $(k-1)(m-k-1)$ -dimensional manifold

$$X^o = \text{Gr}(k, m)^o / (\mathbb{C}^*)^m. \quad (11.11)$$

**Example 11.13** ( $k = 2$ ) For  $k = 2$ , the very affine variety in (11.11) has dimension  $m - 3$ , and it is the moduli space of  $m$  distinct labeled points on the complex projective line  $\mathbb{P}^1$ . This space is ubiquitous in algebraic geometry where it is known as  $\mathcal{M}_{0,m}$ . The punchline of our discussion here is that we interpret the moduli space  $\mathcal{M}_{0,m}$  as a statistical model. And, we then argue that its ML degree is equal to  $(m-3)!$ .

For instance, if  $m = 4$  then  $X^o = \mathcal{M}_{0,4}$  is the Riemann sphere  $\mathbb{P}^1$  with three points removed. The signed Euler characteristic of this surface is one, and Theorem 11.8 applies.

**Proposition 11.14** *The configuration space  $X^o$  in (11.11) is a very affine variety, with coordinates given by the  $k \times k$  minors of the following  $k \times m$  matrix, which we denote by  $M_{k,m}$ :*

$$\begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & (-1)^k & 1 & 1 & 1 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & (-1)^{k-1} & 0 & 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,m-k-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & -1 & \cdots & 0 & 0 & 1 & x_{k-3,1} & x_{k-3,2} & \cdots & x_{k-3,m-k-1} \\ 0 & 1 & 0 & \cdots & 0 & 0 & 1 & x_{k-2,1} & x_{k-2,2} & \cdots & x_{k-2,m-k-1} \\ -1 & 0 & 0 & \cdots & 0 & 0 & 1 & x_{k-1,1} & x_{k-1,2} & \cdots & x_{k-1,m-k-1} \end{bmatrix}.$$

To be precise, the coordinates on  $X^o \subset (\mathbb{C}^*)^{m \choose k}$  are the non-constant minors  $p_{i_1 i_2 \dots i_k}$  of the matrix  $M_k$ .

Following [2, equation (4)], the antidiagonal matrix in the left  $k \times k$  block of  $M_{k,m}$  is chosen so that each unknown  $x_{i,j}$  is precisely equal to  $p_{i_1 i_2 \dots i_k}$  for some  $i_1 < i_2 < \dots < i_k$ . No signs are needed. The scattering potential for the CEGM model is the following multivalued function on  $X^o$ :

$$\ell_u = \sum_{i_1 i_2 \dots i_k} u_{i_1 i_2 \dots i_k} \cdot \log(p_{i_1 i_2 \dots i_k}). \quad (11.12)$$

The critical point equations, known as *scattering equations* [2, equation (7)], are given by

$$\frac{\partial \ell_u}{\partial x_{i,j}} = 0 \quad \text{for } 1 \leq i \leq k-1 \text{ and } 1 \leq j \leq m-k-1. \quad (11.13)$$

These are equations of rational functions. Solving these equations is the agenda in [39, 40, 171].

**Corollary 11.15** *The number of complex solutions to (11.13) is the ML degree of the CEGM model  $X^o$ . This number equals the signed Euler characteristic  $(-1)^{(k-1)(m-k-1)} \cdot \chi(X^o)$ .*

**Example 11.16** ( $k = 2, m = 6$ ) The very affine threefold  $X^o = \mathcal{M}_{0,6}$  is embedded in  $(\mathbb{C}^*)^9$  via

$$\begin{aligned} p_{24} &= x_1, p_{25} = x_2, p_{26} = x_3, p_{34} = x_1 - 1, p_{35} = x_2 - 1, \\ p_{36} &= x_3 - 1, p_{45} = x_2 - x_1, p_{46} = x_3 - x_1, p_{56} = x_3 - x_2. \end{aligned}$$

These nine coordinates on  $X^o \subset (\mathbb{C}^*)^9$  are the non-constant  $2 \times 2$  minors of our matrix

$$M_{2,6} = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 1 \\ -1 & 0 & 1 & x_1 & x_2 & x_3 \end{bmatrix}.$$

The scattering potential is the analogue to the log-likelihood function in statistics:

$$\ell_u = u_{24} \log(p_{24}) + u_{25} \log(p_{25}) + \dots + u_{56} \log(p_{56}).$$

This function has six critical points in  $X^o$ . Hence  $\text{MLdegree}(X^o) = -\chi(X^o) = 6$ .

We now examine the number of critical points of the scattering potential (11.12).

**Theorem 11.17** *The known values of the ML degree for the CEGM model (11.11) are as follows. For  $k = 2$ , the ML degree equals  $(m-3)!$  for all  $m \geq 4$ . For  $k = 3$ , the ML degree equals 2, 26, 1272, 188112, 74570400 when the number of points is  $m = 5, 6, 7, 8, 9$ . For  $k = 4, m = 8$ , the ML degree equals 5211816.*

**Proof** We refer to [2, Example 2.2], [2, Theorem 5.1] and [2, Theorem 6.1] for  $k = 2, 3, 4$ .  $\square$

Knowing these ML degrees helps in solving the scattering equations reliably. It was demonstrated in [2, 171] how this can be done in practice, namely with the software `HomotopyContinuation.jl` [33, 34]. For instance, we see in [171, Table 1] that the  $10! = 3628800$  critical points for  $k = 2, m = 13$  are found in under one hour. See [2, Section 6] for the solution in the challenging case  $k = 4, m = 8$ .

## 11.4 Gaussian Models

We now change topic by turning to models for Gaussian random variables. Let  $\text{PD}_n$  denote the set of positive-definite symmetric  $n \times n$  matrices, i.e. matrices all of whose eigenvalues are positive. This is an open convex cone in a real vector space in the matrix space  $\text{Sym}_2(\mathbb{R}^n)$ , which has dimension  $\binom{n+1}{2}$ . This cone now plays the role which was played by the simplex  $\Delta_n$  when we discussed discrete models.

Given a mean vector  $\mu \in \mathbb{R}^n$  and a covariance matrix  $\Sigma \in \text{PD}_n$ , the associated *Gaussian distribution* is supported on  $\mathbb{R}^n$ . Its density has the familiar “bell shape”; it is the function

$$f_{\mu, \Sigma}(x) := \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} \cdot \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right).$$

We fix a model  $Y \subset \mathbb{R}^n \times \text{PD}_n$  that is defined by polynomial equations in  $(\mu, \Sigma)$ . Suppose we are given  $N$  samples  $U^{(1)}, \dots, U^{(N)}$ . These samples are vectors in  $\mathbb{R}^n$ . They are summarized in the *sample mean*  $\bar{U} = \frac{1}{N} \sum_{i=1}^N U^{(i)}$  and in the *sample covariance matrix*  $S = \frac{1}{N} \sum_{i=1}^N (U^{(i)} - \bar{U})(U^{(i)} - \bar{U})^T$ . Given this representation of the data, the log-likelihood is the following function in the unknowns  $(\mu, \Sigma)$ :

$$\ell(\mu, \Sigma) = -\frac{N}{2} \cdot \left[ \log \det \Sigma + \text{trace}(S \Sigma^{-1}) + (\bar{U} - \mu)^T \Sigma^{-1}(\bar{U} - \mu) \right]. \quad (11.14)$$

The task of likelihood inference is to minimize this function subject to  $(\mu, \Sigma) \in Y$ .

There are two extreme cases. First, consider a model where  $\Sigma$  is fixed to be the identity matrix  $\text{Id}_n$ . Then  $Y = X \times \{\text{Id}_n\}$  and we are supposed to minimize the Euclidean distance from the sample mean  $\bar{U}$  to the variety  $X$  in  $\mathbb{R}^n$ . This is precisely the earlier ED problem.

We instead focus on the second case, the family of *centered Gaussians*, where the mean vector  $\mu$  is fixed to be zero. The model has the form  $\{0\} \times X$ , where  $X$  is a variety in the space  $\text{Sym}_2(\mathbb{R}^n)$  of symmetric  $n \times n$  matrices. Following [173, Proposition 7.1.10], our task is now as follows:

$$\text{Minimize the function } \Sigma \mapsto \log \det \Sigma + \text{trace}(S \Sigma^{-1}) \text{ subject to } \Sigma \in X. \quad (11.15)$$

Using the concentration matrix  $K = \Sigma^{-1}$ , we can write this equivalently as follows:

$$\text{Maximize the function } K \mapsto \log \det K - \text{trace}(SK) \text{ subject to } K \in X^{-1}. \quad (11.16)$$

Here the variety  $X^{-1}$  is the Zariski closure of the set of inverses of all matrices in  $X$ .

**Remark 11.18** The optimization problem (11.15)-(11.16) has a metric interpretation as in (11.4). Namely, we can define the KL divergence between two probability distributions on  $\mathbb{R}^n$  by replacing the sum in (11.2) with the corresponding integral over  $\mathbb{R}^n$ . For two Gaussians we obtain a certain kind of distance between the unknown  $\Sigma$  and the sample covariance matrix  $S$ .

The critical equations of the optimization problem (11.15)-(11.16) can be written as polynomials, since the partial derivatives of the logarithm are rational functions. These equations have finitely many complex solutions. Their number is the *ML degree* of the Gaussian statistical model  $X^{-1}$ .

In the remainder of this section we focus on Gaussian models that are described by linear constraints on either the covariance matrix or its inverse, which is the concentration matrix. Let  $\mathcal{L} \subset \text{Sym}_2(\mathbb{R}^n)$  be a linear space of symmetric matrices (LSSM), whose general element is assumed to be invertible. We are interested in the models  $X^{-1} = \mathcal{L}$  and  $X = \mathcal{L}$ . It is convenient to use primal-dual coordinates  $(\Sigma, K)$  to write the respective critical equations.

**Proposition 11.19** Fix an LSSM  $\mathcal{L}$  and its orthogonal complement  $\mathcal{L}^\perp$  for the inner product  $\langle X, Y \rangle = \text{trace}(XY)$ . The critical equations for the linear concentration model  $X^{-1} = \mathcal{L}$  are

$$K \in \mathcal{L} \text{ and } K\Sigma = \text{Id}_n \text{ and } \Sigma - S \in \mathcal{L}^\perp. \quad (11.17)$$

The critical equations for the linear covariance model  $X = \mathcal{L}$  are

$$\Sigma \in \mathcal{L} \text{ and } K\Sigma = \text{Id}_n \text{ and } KSK - K \in \mathcal{L}^\perp. \quad (11.18)$$

**Proof** This is well-known in statistics. For proofs see [172, Propositions 3.1 and 3.3].  $\square$

The system (11.17) is linear in the unknown matrix  $K$ , whereas the last group of equations in (11.18) is quadratic in  $K$ . The numbers of complex solutions are the *ML degree* of  $\mathcal{L}$  and the *reciprocal ML degree* of  $\mathcal{L}$ . The former is smaller than the latter, and (11.17) is easier to solve than (11.18).

**Example 11.20** Let  $n = 4$  and  $\mathcal{L}$  a generic LSSM of dimension  $k$ . Our degrees are as follows:

|                           |   |    |    |    |    |    |    |   |
|---------------------------|---|----|----|----|----|----|----|---|
| $k = \dim(\mathcal{L}) :$ | 2 | 3  | 4  | 5  | 6  | 7  | 8  | 9 |
| ML degree :               | 3 | 9  | 17 | 21 | 21 | 17 | 9  | 3 |
| reciprocal ML degree :    | 5 | 19 | 45 | 71 | 81 | 63 | 29 | 7 |

These numbers and many more appear in [172, Table 1].

ML degrees and reciprocal ML degrees of linear spaces of symmetric matrices have been studied intensively in the recent literature, both for generic and special spaces  $\mathcal{L}$ . See [6, 23, 71] and the references therein. We now present an important result due to Manivel, Michalek, Monin, Seynnaeve, Vodička and Wiśniewski. Theorem 11.21 paraphrases highlights from their articles [129, 134].

**Theorem 11.21** The ML degree of a generic linear subspace  $\mathcal{L}$  of dimension  $k$  in  $\text{Sym}_2(\mathbb{R}^n)$  is the number of quadrics in  $\mathbb{P}^{n-1}$  that pass through  $\binom{n+1}{2} - k$  general points and are tangent to  $k - 1$  general hyperplanes. For fixed  $k$ , this number is a polynomial in  $n$  of degree  $k - 1$ .

**Proof** The first statement is [134, Corollary 2.6 (4)], here interpreted classically in terms of Schubert calculus. For a detailed discussion see the introduction of [129]. The second statement appears in [129, Theorem 1.3 and Corollary 4.13].  $\square$

**Example 11.22** ( $n = 4$ ) Fix  $10 - k$  points and  $k - 1$  planes in  $\mathbb{P}^3$ . We are interested in all quadratic surfaces that contain the points and are tangent to the planes. This points and planes impose 9 constraints on  $\mathbb{P}(\text{Sym}_2(\mathbb{C}^4)) \simeq \mathbb{P}^9$ . Passing through a point is a linear equation. Being tangent to a plane is a cubic constraint on  $\mathbb{P}^9$ . Bézout's Theorem suggests that there could be  $3^{k-1}$  solutions. This is correct for  $k \leq 3$  but it overcounts for  $k \geq 4$ . Indeed, in Example 11.20 we see 17, 21, 21, ... instead of 27, 81, 243, ...

The intersection theory approach in [129, 134] leads to formulas for the ML degrees of linear Gaussian models. From this we obtain provably correct numerical methods for maximum likelihood estimation. Namely, after computing critical points as in [172], we can certify them as in [33]. Since the ML degree is known, one can then be sure that all solutions have been found.



## **Chapter 12**

### **Tensors**

Tensors are generalizations of matrices. They can be viewed as tables of higher dimension. A  $2 \times 2$ -matrix is a table that contains 4 numbers aligned in two directions and each direction has dimension 2; a  $2 \times 2 \times 2$  tensor is a table with 8 numbers aligned in three directions where each direction has dimension 2.

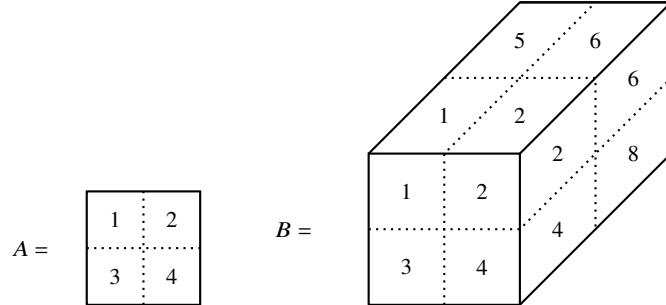


Fig. 12.1: A  $2 \times 2$  matrix  $A$  and a  $2 \times 2 \times 2$  tensor  $B$ .

In this chapter we offer an introduction to tensors, with a view towards metric algebraic geometry. We present the notion of tensor rank, along with the Segre variety and the Veronese variety. These arise from tensors of bounded rank. We then turn to the spectral theory of tensors, and we show that eigenvectors and singular vectors arise naturally in the study of Euclidean distance problems. In the last section we turn to volumes in complex projective space, and we determine these volumes for Segre and Veronese varieties.

## 12.1 Tensors and their Rank

For positive integers  $n_1, \dots, n_d$ , the vector space of real  $n_1 \times \dots \times n_d$ -tensors is denoted by

$$\mathbb{R}^{n_1 \times \dots \times n_d} := \{ A = (a_{i_1, \dots, i_d})_{1 \leq i_1 \leq n_1, \dots, 1 \leq i_d \leq n_d} \mid a_{i_1, \dots, i_d} \in \mathbb{R} \}.$$

In the literature, elements in  $\mathbb{R}^{n_1 \times \dots \times n_d}$  are also sometimes called *hypermatrices* or *Cartesian tensors*, because our definition of a tensor as a multidimensional array relies on a basis. The dimension of the vector space  $\mathbb{R}^{n_1 \times \dots \times n_d}$  is  $n_1 \cdots n_d$ . We also consider the space of complex tensors  $\mathbb{C}^{n_1 \times \dots \times n_d}$ . The number  $d$  is called the *order* of the tensor. Tensors of order one are vectors. Tensors of order two are matrices.

A common notation for order-three tensors is by writing them as a tuple of matrices.

**Example 12.1** The tensor  $B = (b_{ijk})_{1 \leq i,j,k \leq 2}$  from Figure 12.1 can be written as

$$B = \left( \begin{array}{cc|cc} 1 & 2 & 5 & 6 \\ 3 & 4 & 7 & 8 \end{array} \right) \in \mathbb{R}^{2 \times 2 \times 2}.$$

The left matrix contains the entries  $b_{ij1}$  and the matrix on the right contains the entries  $b_{ij2}$ . ◊

The Euclidean inner product of two tensors  $A = (a_{i_1, \dots, i_d})$  and  $B = (b_{i_1, \dots, i_d})$  is defined as

$$\langle A, B \rangle := \sum_{i_1=1}^{n_1} \cdots \sum_{i_d=1}^{n_d} a_{i_1, \dots, i_d} \cdot b_{i_1, \dots, i_d}$$

The resulting norm on  $\mathbb{R}^{n_1 \times \dots \times n_d}$  is given by the familiar formula  $\|A\| = \sqrt{\langle A, A \rangle}$ .

One way to create a tensor is to take the *outer product* of  $d$  vectors  $\mathbf{v}_i \in \mathbb{R}^{n_i}$ , where  $1 \leq i \leq d$ . This is defined by taking all  $n_1 n_2 \cdots n_d$  possible products between the entries of distinct vectors:

$$\mathbf{v}_1 \otimes \cdots \otimes \mathbf{v}_d := \left( (\mathbf{v}_1)_{i_1} \cdots (\mathbf{v}_d)_{i_d} \right)_{1 \leq i_1 \leq n_1, \dots, 1 \leq i_d \leq n_d}. \quad (12.1)$$

The outer products of the standard basis vectors  $\mathbf{e}_i$  give the standard basis of  $\mathbb{R}^{n_1 \times \cdots \times n_d}$ . Thus, we have

$$(a_{i_1, \dots, i_d}) = \sum_{i_1=1}^{n_1} \cdots \sum_{i_d=1}^{n_d} a_{i_1, \dots, i_d} \mathbf{e}_{i_1} \otimes \cdots \otimes \mathbf{e}_{i_d}.$$

If  $n := n_1 = \cdots = n_d$  are all equal, then we write the tensor space as  $(\mathbb{R}^n)^{\otimes d} := \mathbb{R}^{n \times \cdots \times n}$ . Each permutation  $\pi \in \mathfrak{S}_d$  acts on outer products in  $(\mathbb{R}^n)^{\otimes d}$  via  $\pi(\mathbf{v}_1 \otimes \cdots \otimes \mathbf{v}_d) = \mathbf{v}_{\pi(1)} \otimes \cdots \otimes \mathbf{v}_{\pi(d)}$ . This yields an action of the symmetric group  $\mathfrak{S}_d$  on  $(\mathbb{R}^n)^{\otimes d}$  via linear extension. For  $d = 2$ , this is transposition of matrices.

**Definition 12.2** A tensor  $A \in (\mathbb{R}^n)^{\otimes d}$  is called *symmetric* if  $\pi(A) = A$  for all  $\pi \in \mathfrak{S}_d$ . We abbreviate

$$S^d(\mathbb{R}^n) := \{A \in (\mathbb{R}^n)^{\otimes d} \mid A \text{ is symmetric}\}.$$

The space  $(\mathbb{R}^n)^{\otimes d}$  of symmetric tensors is the image of the linear map  $A \mapsto \sum_{\pi \in \mathfrak{S}_d} \pi(A)$ , so it is a linear subspace of  $(\mathbb{R}^n)^{\otimes d}$ . Its dimension is  $\binom{n+d-1}{d}$ . For  $\mathbf{v} \in \mathbb{R}^n$ , we write the  $d$ -fold outer product as

$$\mathbf{v}^{\otimes d} := \mathbf{v} \otimes \cdots \otimes \mathbf{v} \in S^d(\mathbb{R}^n).$$

Every symmetric tensor  $A \in S^d(\mathbb{R}^n)$  determines a homogeneous polynomial of degree  $d$  in  $n$  variables:

$$F_A(\mathbf{x}) := \sum_{i_1=1}^n \sum_{i_2=1}^n \cdots \sum_{i_d=1}^n a_{i_1, i_2, \dots, i_d} x_{i_1} x_{i_2} \cdots x_{i_d}. \quad (12.2)$$

The map  $A \mapsto F_A$  is a isomorphism between  $S^d(\mathbb{R}^n)$  and the vector space of homogeneous polynomials of degree  $d$  in  $n$  variables. Their common dimension is  $\binom{n+d-1}{d}$ . We can write the polynomial in (12.2) as

$$F_A(\mathbf{x}) = \langle A, \mathbf{x}^{\otimes d} \rangle.$$

For  $A \in S^d(\mathbb{R}^n)$ , we have  $\|A\| = \|F_A\|_{\text{BW}}$ , where the latter is the Bombieri-Weyl metric from Chapter 9.

**Example 12.3** Symmetric tensors of order three are in bijection with homogeneous cubics. For instance,

$$\text{For } A = \begin{pmatrix} 1 & -1 \\ -1 & 0 \end{pmatrix} \in S^3(\mathbb{R}^2) \text{ we have } F_A(\mathbf{x}) = x_1^3 - 3x_1^2 x_2 + 2x_2^3.$$

This generalizes the familiar bijection between symmetric matrices and quadratic forms.  $\diamond$

Given a tuple of matrices  $(M_1, \dots, M_d) \in \mathbb{R}^{k_1 \times n_1} \times \cdots \times \mathbb{R}^{k_d \times n_d}$ , we define the *multilinear multiplication*

$$(M_1, \dots, M_d).(\mathbf{v}_1 \otimes \cdots \otimes \mathbf{v}_d) := (M_1 \mathbf{v}_1) \otimes \cdots \otimes (M_d \mathbf{v}_d).$$

This extends by linearity to all of  $\mathbb{R}^{n_1 \times \cdots \times n_d}$ . For  $k_1 = n_1, \dots, k_d = n_d$ , this induces a representation of  $\text{GL}(n_1) \times \cdots \times \text{GL}(n_d)$  into the general linear group of  $\mathbb{R}^{n_1 \times \cdots \times n_d}$ . If  $d = 2$  then  $(M_1, M_2).A = M_1 A M_2^\top$ . Thus, multilinear multiplication is a generalization of simultaneous left-right multiplication for matrices.

A central topic of this chapter are *rank-one tensors*. A tensor  $A$  has rank one if has the form (12.1). For  $d = 2$ , this gives us  $n_1 \times n_2$  matrices of rank one. Note that the inner product between rank-one tensors is

$$\langle \mathbf{v}_1 \otimes \cdots \otimes \mathbf{v}_d, \mathbf{w}_1 \otimes \cdots \otimes \mathbf{w}_d \rangle = \langle \mathbf{v}_1, \mathbf{w}_1 \rangle \cdots \langle \mathbf{v}_d, \mathbf{w}_d \rangle. \quad (12.3)$$

**Definition 12.4** The (real) *Segre variety*  $\mathcal{S}_{n_1, \dots, n_d}$  consists of all  $n_1 \times \dots \times n_d$ -tensors of rank one, i.e.

$$\mathcal{S}_{n_1, \dots, n_d} := \{\mathbf{v}_1 \otimes \dots \otimes \mathbf{v}_d \mid \mathbf{v}_i \in \mathbb{R}^{n_i}, 1 \leq i \leq d\}.$$

We simply write  $\mathcal{S}$  when  $n_1, \dots, n_d$  are clear from the context. The most immediate way to see that  $\mathcal{S}$  is an algebraic variety goes as follows. We can always *flatten* a tensor  $A \in \mathbb{R}^{n_1 \times \dots \times n_d}$  into  $d$  matrices  $F_1, \dots, F_d$ , where  $F_i \in \mathbb{R}^{n_i \times (\prod_{j \neq i} n_j)}$ . Then, we have  $A \in \mathcal{S}$  if and only if the column span of  $F_i$  has dimension at most one for  $1 \leq i \leq d$ . This is equivalent to saying that the  $F_i$  are all of rank at most 1; i.e., their  $2 \times 2$ -minors vanish. The dimension of the Segre variety is  $\dim \mathcal{S} = n_1 + \dots + n_d + 1 - d$ .

The analogue of the Segre variety for symmetric tensors is the Veronese variety.

**Definition 12.5** The *Veronese variety* is the variety of symmetric tensors of rank one. In symbols,

$$\mathcal{V} := \mathcal{V}_{n,d} := \{\mathbf{v}^{\otimes d} \mid \mathbf{v} \in \mathbb{R}^n\}.$$

This is the intersection of  $\mathcal{S}$  with a linear subspace. Hence  $\mathcal{V}$  is an algebraic variety. Note that  $\dim \mathcal{V} = n$ .

**Example 12.6** Let  $\mathbf{v} = (x, y) \in \mathbb{R}^2$ . Then,  $\mathbf{v}^{\otimes 3}$  comprises all monomials of degree three in  $x$  and  $y$ :

$$\mathbf{v}^{\otimes 3} = \begin{pmatrix} x^3 & x^2y & x^2y & xy^2 \\ x^2y & xy^2 & xy^2 & y^3 \end{pmatrix}.$$

Here, the surface  $\mathcal{V}$  is the cone over the twisted cubic curve. In general, the distinct entries of  $\mathbf{v}^{\otimes d}$  are all monomials of degree  $d$  in the entries of  $\mathbf{v}$ . The Veronese variety  $\mathcal{V}$  is parametrized by these monomials.  $\diamond$

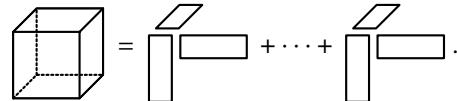
The Segre variety induces a notion of rank of a tensor  $A \in \mathbb{R}^{n_1 \times \dots \times n_d}$ . Namely, we define

$$\text{rank}(A) := \min \{r \geq 0 \mid \text{there exists } A_1, \dots, A_r \in \mathcal{S} \text{ with } A = A_1 + \dots + A_r\}.$$

For matrices ( $d = 2$ ), this coincides with the usual matrix rank. We denote tensors of rank at most  $r$  by

$$\Sigma_r := \Sigma_{r, n_1, \dots, n_d} := \{A \in \mathbb{R}^{n_1 \times \dots \times n_d} \mid \text{rank}(A) \leq r\}.$$

A tensor of order three and rank  $r$  can be visualized as follows:



In the case of matrices, the set  $\Sigma_r$  is a variety, defined by the  $(r+1) \times (r+1)$ -minors. For  $d \geq 3$  and  $r \geq 2$ , however,  $\Sigma_r$  is no longer a variety. This is implied by the following result due to da Silva and Lim [55].

**Proposition 12.7** The set  $\Sigma_2$  of tensors of rank  $\leq 2$  is generally not closed in the Euclidean topology.

**Proof** Fix  $d = 3$  and  $n_1 = n_2 = n_3 = 2$ . Consider linearly independent pairs of vectors  $\{\mathbf{x}_1, \mathbf{x}_2\}$ ,  $\{\mathbf{y}_1, \mathbf{y}_2\}$ , and  $\{\mathbf{z}_1, \mathbf{z}_2\}$  in  $\mathbb{R}^2$ . For any  $\varepsilon > 0$ , we define a rank-two tensor  $A_\varepsilon \in \Sigma_2$  as follows:

$$A_\varepsilon := \varepsilon^{-1} (\mathbf{x}_1 + \varepsilon \mathbf{x}_2) \otimes (\mathbf{y}_1 + \varepsilon \mathbf{y}_2) \otimes (\mathbf{z}_1 + \varepsilon \mathbf{z}_2) - \varepsilon^{-1} \mathbf{x}_1 \otimes \mathbf{y}_1 \otimes \mathbf{z}_1. \quad (12.4)$$

Then,  $A := \lim_{\varepsilon \rightarrow 0} A_\varepsilon = \mathbf{x}_1 \otimes \mathbf{y}_1 \otimes \mathbf{z}_2 + \mathbf{x}_1 \otimes \mathbf{y}_2 \otimes \mathbf{z}_1 + \mathbf{x}_2 \otimes \mathbf{y}_1 \otimes \mathbf{z}_1$ . We claim that  $A$  has rank three. For  $\mathbf{h} \in \mathbb{R}^2$ , we apply multilinear multiplication by  $X := (\mathbf{I}_2, \mathbf{I}_2, \mathbf{h}^\top)$  to  $A$ . This gives the  $2 \times 2$ -matrix

$$X \cdot A = \langle \mathbf{h}, \mathbf{z}_2 \rangle \cdot \mathbf{x}_1 \otimes \mathbf{y}_1 + \langle \mathbf{h}, \mathbf{z}_1 \rangle \cdot \mathbf{x}_1 \otimes \mathbf{y}_2 + \langle \mathbf{h}, \mathbf{z}_1 \rangle \cdot \mathbf{x}_2 \otimes \mathbf{y}_1. \quad (12.5)$$

Choosing  $\mathbf{h}$  with  $\langle \mathbf{h}, \mathbf{z}_1 \rangle \neq 0$  yields a matrix of rank two, and hence  $A$  has rank at least two. We suppose now that  $A = \mathbf{u}_1 \otimes \mathbf{v}_1 \otimes \mathbf{w}_1 + \mathbf{u}_2 \otimes \mathbf{v}_2 \otimes \mathbf{w}_2$  has rank two. Among the three pairs of vectors, at least two pairs are linearly independent. Suppose the pairs  $(\mathbf{u}_1, \mathbf{u}_2)$  and  $(\mathbf{v}_1, \mathbf{v}_2)$  are linearly independent. Then

$$X.A = \langle \mathbf{h}, \mathbf{w}_1 \rangle \cdot \mathbf{u}_1 \otimes \mathbf{v}_1 + \langle \mathbf{h}, \mathbf{w}_2 \rangle \cdot \mathbf{u}_2 \otimes \mathbf{v}_2. \quad (12.6)$$

Let us now pick a non-zero  $\mathbf{h}$  such that  $\langle \mathbf{h}, \mathbf{z}_1 \rangle = 0$ . By (12.5),  $X.A$  has rank one and hence  $\langle \mathbf{h}, \mathbf{w}_1 \rangle = 0$  or  $\langle \mathbf{h}, \mathbf{w}_2 \rangle = 0$  by (12.6). We may assume  $\langle \mathbf{h}, \mathbf{w}_1 \rangle = 0$ . This implies that  $\mathbf{w}_1$  is a multiple of  $\mathbf{z}_1$ , since both are perpendicular to  $\mathbf{h}$ . First, if the pair  $(\mathbf{w}_1, \mathbf{w}_2)$  is linearly independent, there is another  $\mathbf{h}'$  with  $\langle \mathbf{h}', \mathbf{w}_1 \rangle \neq 0$  but  $\langle \mathbf{h}', \mathbf{w}_2 \rangle = 0$ . Then,  $\langle \mathbf{h}', \mathbf{z}_1 \rangle \neq 0$  and so  $X.A$  has rank two by (12.5) and it has rank one by (12.6); a contradiction. Second, if  $\mathbf{w}_1, \mathbf{w}_2$  and  $\mathbf{z}_1$  are multiples of one another, then  $A$  is of the form  $A = (\lambda_1 \mathbf{u}_1 \otimes \mathbf{v}_1 + \lambda_2 \mathbf{u}_2 \otimes \mathbf{v}_2) \otimes \mathbf{z}_1$  (i.e., its “third row space” is spanned by  $\mathbf{z}_1$ ). This is a contradiction to  $A = \mathbf{x}_1 \otimes \mathbf{y}_1 \otimes \mathbf{z}_2 + (\mathbf{x}_1 \otimes \mathbf{y}_2 \otimes \mathbf{z}_2) \otimes \mathbf{z}_1$ . In both cases, we have derived a contradiction, so  $A$  cannot have rank two, but it must have rank three. Finally, the same proof extends to larger tensor formats. Namely, we can simply embed  $A$  into a larger tensor format while retaining the relevant rank properties.  $\square$

A decomposition of the form  $A = A_1 + \cdots + A_r$  with  $A_i \in \mathcal{S}$  is called a *rank- $r$  decomposition*. In the signal processing literature it is also called *canonical polyadic decomposition*. One appealing property of higher order tensors is *identifiability*. That is, many tensor decompositions are actually unique (rank decompositions of matrices are never unique). The following is [50, Theorem 1.1] and [153, Lemma 28].

**Theorem 12.8** Let  $d \geq 3$  and  $n_1 \geq \cdots \geq n_d$  with  $\prod_{i=1}^d n_i \leq 15000$ , and set

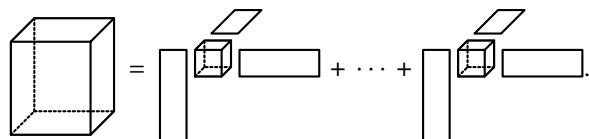
$$r_0 = \left\lceil \frac{\dim \mathbb{R}^{n_1 \times \cdots \times n_d}}{\dim \mathcal{S}} \right\rceil = \left\lceil \frac{n_1 \cdots n_d}{1 + \sum_{i=1}^d (n_i - 1)} \right\rceil.$$

Suppose that  $r < r_0$  and the tuple  $(n_1, \dots, n_d, r)$  is not among the following special cases:

| $(n_1, \dots, n_d)$                                | $r$   |
|--|---|
| $(4, 4, 3)$  | 5   |
| $(4, 4, 4)$  | 6   |
| $(6, 6, 3)$  | 8   |
| $(n, n, 2, 2)$                                     | $2n - 1$  |
| $(2, 2, 2, 2, 2)$                                  | 5   |
| $n_1 > \prod_{i=2}^d n_i - \sum_{i=2}^d (n_i - 1)$ | $r \geq \prod_{i=2}^d n_i - \sum_{i=2}^d (n_i - 1)$ |

Then, a general tensor in  $\Sigma_r$  has a unique rank- $r$  decomposition.

**Remark 12.9** There are other notions of tensor rank, and these give alternative decompositions. Let  $\mathbf{k} = (k_1, \dots, k_d)$  be a vector of integers with  $1 \leq k_i \leq n_i$ . Let  $A \in \mathbb{R}^{n_1 \times \cdots \times n_d}$  and  $F_1, \dots, F_d$  be the flattenings of  $A$ . Then, we say that  $A$  has *multilinear rank* (at most)  $\mathbf{k}$  if  $\text{rank}(F_j) \leq k_j$  for all  $1 \leq j \leq d$ . A *block term decomposition* of the tensor  $A$  has the form  $A = A_1 + \cdots + A_r$ , where each  $A_i$  has low multilinear rank  $\mathbf{k}$ . For order-three tensors, a block term decomposition can be visualized as follows:



The rank decomposition is the special case  $\mathbf{k} = (1, \dots, 1)$ . In statistics, this corresponds to mixtures of independence models. Block term decompositions represent mixtures of discrete probability distributions that are more complex. Identifiability for block term decompositions is less studied than for rank decompositions. Some results, primarily for the case  $d = 3$ , can be found in [121] and references therein.  $\diamond$

## 12.2 Eigenvectors and Singular Vectors

In this section we study the Euclidean distance problem for the Segre variety and the Veronese variety. We start with the observation that each of them is the image of a polynomial map. The two maps are

$$\begin{aligned}\psi : \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_d} &\rightarrow \mathcal{S}, (\mathbf{v}_1, \dots, \mathbf{v}_d) \mapsto \mathbf{v}_1 \otimes \cdots \otimes \mathbf{v}_d, \\ v : \mathbb{R}^n &\rightarrow \mathcal{V}, \quad \mathbf{v} \mapsto \mathbf{v}^{\otimes d}.\end{aligned}\tag{12.7}$$

We first consider the ED problem for the Segre variety  $\mathcal{S}$ . Given any tensor  $A \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ , we seek the rank one tensor  $B$  which is closest to  $A$  in Euclidean norm. Thus, we must solve the optimization problem

$$\min_{B \in \mathcal{S}} \|A - B\| = \min_{\mathbf{v}_i \in \mathbb{R}^{n_i}} \|A - \psi(\mathbf{v}_1, \dots, \mathbf{v}_d)\|. \tag{12.8}$$

*Remark 12.10* Consider the best rank  $r$  approximation for a tensor  $A$  of order  $d \geq 3$ . This means computing  $\min_{B \in \Sigma_r} \|A - B\|$ . For  $r = 1$  this is precisely (12.8) and we shall present the solution in this section. For  $r \geq 2$ , this problem can be ill-posed (i.e no minimizer exists). In fact, Proposition 12.7 implies that this problem is ill-posed for all tensors  $A$  in a full-dimensional open subset of  $\mathbb{R}^{n_1 \times \cdots \times n_d}$ . This was first noted by da Silva and Lim [55]. This is different for matrices, where the Eckart-Young Theorem (Theorem 2.10) provides an explicit algorithm for computing the minimizer. Likewise, for  $r = 1$ , the image of the Segre map  $\psi$  is closed, and hence the problem (12.8) has a solution  $B \in \mathcal{S}$ . Our goal is to identify this  $B$ .

The critical values of the distance function  $\mathcal{S} \rightarrow \mathbb{R}, B \mapsto \|A - B\|$  are in one-to-one correspondence to the critical values of  $\|\psi(\mathbf{v}_1, \dots, \mathbf{v}_d)\|$ , up to scaling  $(\mathbf{v}_1, \dots, \mathbf{v}_d) \mapsto (t_1 \mathbf{v}_1, \dots, t_d \mathbf{v}_d)$  with  $t_1 \cdots t_d = 1$ . Our goal is therefore to find the critical points of the function that maps the tuple  $(\mathbf{v}_1, \dots, \mathbf{v}_d)$  to

$$\|A - \psi(\mathbf{v}_1, \dots, \mathbf{v}_d)\|^2 = \|A\|^2 - 2\langle A, \psi(\mathbf{v}_1, \dots, \mathbf{v}_d) \rangle + \|\psi(\mathbf{v}_1, \dots, \mathbf{v}_d)\|^2. \tag{12.9}$$

By (12.3), we have  $\|\psi(\mathbf{v}_1, \dots, \mathbf{v}_d)\|^2 = \|\mathbf{v}_1\|^2 \cdots \|\mathbf{v}_d\|^2$ . The gradient of (12.9) with respect to  $\mathbf{v}_k$  equals

$$\begin{aligned}& -2 \frac{d}{d\mathbf{v}_k} \langle A, \psi(\mathbf{v}_1, \dots, \mathbf{v}_d) \rangle + \frac{d}{d\mathbf{v}_k} (\|\mathbf{v}_1\|^2 \cdots \|\mathbf{v}_d\|^2) \\&= -2 \left( \begin{array}{c} \langle A, \mathbf{v}_1 \otimes \cdots \otimes \mathbf{v}_{k-1} \otimes \mathbf{e}_1 \otimes \mathbf{v}_{k+1} \otimes \cdots \otimes \mathbf{v}_d \rangle \\ \vdots \\ \langle A, \mathbf{v}_1 \otimes \cdots \otimes \mathbf{v}_{k-1} \otimes \mathbf{e}_{n_k} \otimes \mathbf{v}_{k+1} \otimes \cdots \otimes \mathbf{v}_d \rangle \end{array} \right) + \left( 2 \prod_{j \neq k} \|\mathbf{v}_j\|^2 \right) \mathbf{v}_k.\end{aligned}\tag{12.10}$$

Here  $\mathbf{e}_1, \dots, \mathbf{e}_{n_k}$  denotes the standard basis of  $\mathbb{R}^{n_k}$ . We abbreviate  $\sigma_k := \prod_{j \neq k} \|\mathbf{v}_j\|^2$ . With this convention, we write the critical equations  $\frac{d}{d\mathbf{v}_k} \|A - \psi(\mathbf{v}_1, \dots, \mathbf{v}_d)\|^2 = 0$  from (12.10) in the following compact form:

$$A \bullet \otimes_{j \neq k} \mathbf{v}_j = \sigma_k \cdot \mathbf{v}_k \quad \text{for } k = 1, \dots, d. \tag{12.11}$$

For  $d = 3$ , we can visualize this as follows

$$A \bullet \otimes_{j \neq k} \mathbf{v}_j = \sigma_k \cdot \mathbf{v}_k \tag{12.12}$$

**Definition 12.11** Let  $A \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ . We say that  $(\mathbf{v}_1, \dots, \mathbf{v}_d) \in \mathbb{C}^{n_1} \times \cdots \times \mathbb{C}^{n_d}$ ,  $\mathbf{v}_i \neq 0$ , is a *singular vector tuple* for  $A$  if (12.11) holds for some  $\sigma_1, \dots, \sigma_d \in \mathbb{C}$ . If this holds then  $(t_1 \mathbf{v}_1, \dots, t_d \mathbf{v}_d)$  satisfies the same condition, for all  $t_1, \dots, t_d \in \mathbb{C}$ . Thus, singular vector tuples are considered only up scaling.

For  $d = 2$ , this coincides with the classic definition of singular vector pairs for rectangular matrices. The singular value decomposition implies that for a general matrix  $A \in \mathbb{R}^{n_1 \times n_2}$  there are precisely  $\min\{n_1, n_2\}$  singular vector pairs that are all real. For higher order tensors not all singular vectors need to be real.

**Proposition 12.12** *The critical points (over  $\mathbb{C}$ ) of the ED problem for the Segre variety  $\mathcal{S}$  are the tuples of singular vectors. Hence the Euclidean distance degree of  $\mathcal{S}$  equals the number of singular vector tuples.*

**Proof** The computation in (12.10) shows that every critical point of the optimization problem (12.8) is a singular vector tuple of the tensor  $A$ . Conversely, given any solution to the equations (12.11), we can rescale the vectors  $\mathbf{v}_i$  in the tuple so that  $\sigma_k = \prod_{j \neq k} \|\mathbf{v}_j\|^2$  holds for all  $k$ . Hence the critical points are in bijection with the singular vector tuples, assuming that these are viewed as points in  $\mathbb{P}^{n_1-1} \times \cdots \times \mathbb{P}^{n_d-1}$ .  $\square$

**Example 12.13** ( $d = 3$ ) Fix an  $n_1 \times n_2 \times n_3$  tensor  $A = (a_{ijk})$ . Our unknowns are points  $\mathbf{v}_1 = (x_1, \dots, x_{n_1}) \in \mathbb{P}^{n_1-1}$ ,  $\mathbf{v}_2 = (y_1, \dots, y_{n_2}) \in \mathbb{P}^{n_2-1}$ ,  $\mathbf{v}_3 = (z_1, \dots, z_{n_3}) \in \mathbb{P}^{n_3-1}$ . The singular vector equations (12.11) are

$$\begin{aligned}\sum_{j=1}^{n_2} \sum_{k=1}^{n_3} a_{ijk} y_j z_k &= \sigma_1 x_i && \text{for } i = 1, 2, \dots, n_1, \\ \sum_{i=1}^{n_1} \sum_{k=1}^{n_3} a_{ijk} x_i z_k &= \sigma_2 y_j && \text{for } j = 1, 2, \dots, n_2, \\ \sum_{i=1}^{n_1} \sum_{k=1}^{n_2} a_{ijk} x_i y_j &= \sigma_3 z_k && \text{for } k = 1, 2, \dots, n_3.\end{aligned}$$

To work modulo scaling, we set  $x_1 = y_1 = z_1 = 1$ . Then this is a square system of polynomial equations, and it has finitely many solutions for generic tensors  $A = (a_{ijk})$ . For instance, the number of solutions is 6 when  $n_1 = n_2 = n_3 = 2$  and it is 37 when  $n_1 = n_2 = n_3 = 3$ . This will be explained by Theorem 12.14.

To obtain the critical points of our ED problem (12.8), we retain  $x_1, y_1, z_1$  as unknowns, and we add the constraints  $\sigma_k = \prod_{j \neq k} \|\mathbf{v}_j\|^2$  for  $k = 1, 2, 3$ . Explicitly, these constraints are the polynomial equations

$$(x_1^2 + x_2^2)(y_1^2 + y_2^2) = \sigma_3, \quad (x_1^2 + x_2^2)(z_1^2 + z_2^2) = \sigma_2 \quad \text{and} \quad (y_1^2 + y_2^2)(z_1^2 + z_2^2) = \sigma_1.$$

For each singular vector triple, we obtain a curve of solutions that maps to the same critical tensor  $B$  of rank one, up to scaling. For instance, for  $n_1 = n_2 = n_3 = 2$ , there are six such curves. Each of them has degree 12. In other words, the formulation on the right of (12.8) requires some care because the rank one factorization (12.1) of is not unique. By using singular vector tuples up to scaling, we are on the safe side.

The following theorem due to Friedland and Ottaviani [72] gives a combinatorial formula, in terms of  $n_1, n_2, \dots, n_d$ , for number of singular vector tuples, and hence for the ED degree of the Segre variety  $\mathcal{S}$ .

**Theorem 12.14** *Let  $A \in \mathbb{R}^{n_1 \times \cdots \times n_d}$  be a general tensor. The number of singular vector tuples of  $A$ , counted up to scaling  $(\mathbf{v}_1, \dots, \mathbf{v}_d) \mapsto (t_1 \mathbf{v}_1, \dots, t_d \mathbf{v}_d)$ , is the coefficient of  $x_1^{n_1-1} \cdots x_d^{n_d-1}$  in the polynomial*

$$\prod_{i=1}^d \frac{f_i^{n_i} - x_i^{n_i}}{f_i - x_i}, \quad \text{where } f_i := \sum_{j \neq i} x_j.$$

**Example 12.15** We consider the formula in Theorem 12.14 for binary tensors  $A \in \mathbb{R}^{2 \times \cdots \times 2}$ . In this case,

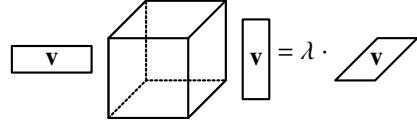
$$\prod_{i=1}^d \frac{f_i^2 - x_i^2}{f_i - x_i} = \prod_{i=1}^d (f_i + x_i) = (x_1 + x_2 + \cdots + x_d)^d.$$

The coefficient of  $x_1 \cdots x_d$  equals  $d!$ . Hence, the ED degree of the Segre variety in  $\mathbb{R}^{2 \times 2 \times 2}$  is  $3! = 6$ .  $\diamond$

We turn pass from general tensors to symmetric tensors, and we consider the ED problem for the Veronese variety  $\mathcal{V}$  in Definition 12.5. Now  $A$  is a symmetric tensor, all  $\mathbf{v}_i$  are equal to the same vector  $\mathbf{v}$  in  $\mathbb{R}^n$ , and the gradient in (12.10) does not depend on  $k$ . Setting  $\lambda := \sigma_k$ , the critical equations for  $\mathcal{V}$  are

$$A \bullet \mathbf{v}^{\otimes(d-1)} = \lambda \mathbf{v}. \quad (12.13)$$

Similarly to (12.12), we visualize this for symmetric tensors of order  $d = 3$ :



**Definition 12.16** Let  $A \in S^d(\mathbb{R}^n)$ . We call  $\mathbf{v} \in \mathbb{C}^n \setminus \{0\}$  an *eigenvector* of  $A$  if the equation (12.13) holds for some  $\lambda \in \mathbb{C}$ . The pair  $(\mathbf{v}, \lambda)$  is called an *eigenpair* of the symmetric tensor  $A$ .

Eigenpairs have another interesting interpretation, next to being critical points for the Euclidean distance problem on the Veronese variety. Recall that every symmetric tensor  $A \in S^d(\mathbb{R}^n)$  corresponds uniquely to a homogeneous polynomial  $F_A(\mathbf{x}) = \langle A, \mathbf{x}^{\otimes d} \rangle$  of degree  $d$  in  $n$  variables. Under this interpretation, the vector  $A \bullet \mathbf{x}^{\otimes(d-1)}$  on the left in (12.13) is the gradient vector  $\nabla_{\mathbf{x}} F_A = (\partial F_A / \partial x_1, \dots, \partial F_A / \partial x_n)$ . Therefore, eigenpairs correspond to fixed points of the gradient map  $\mathbb{P}^{n-1} \dashrightarrow \mathbb{P}^{n-1}, \mathbf{x} \mapsto \nabla_{\mathbf{x}} F_A$ .

**Example 12.17** ( $n = 3$ ) Symmetric tensors of order  $d$  and format  $3 \times 3 \times \dots \times 3$  correspond to ternary forms  $F(x_1, x_2, x_3)$  of degree  $d$ . The eigenvectors of such a tensor are characterized by the equations

$$\partial F / \partial x_1 = \lambda x_1, \quad \partial F / \partial x_2 = \lambda x_2, \quad \partial F / \partial x_3 = \lambda x_3.$$

By eliminating  $\lambda$ , this is equivalent to the statement that the following matrix has rank one:

$$\begin{pmatrix} x_1 & x_2 & x_3 \\ \partial F / \partial x_1 & \partial F / \partial x_2 & \partial F / \partial x_3 \end{pmatrix}$$

This is a dependent system of three homogeneous equations of degree  $d$  in three variables. The number of solutions  $(x_1 : x_2 : x_3)$  in  $\mathbb{P}^2$  equals  $d^2 - d + 1$ . For  $d = 2$ , these are just the three eigenvectors of a symmetric  $3 \times 3$  matrix. For  $d = 3$ , one interesting example is  $F = x_1^3 + x_2^3 + x_3^3$ . The corresponding  $3 \times 3 \times 3$  tensor has 7 eigenvectors, namely the elements of  $\{0, 1\}^3 \setminus \{0\}^3$ . How about  $F = x_1^d + x_2^d + x_3^d$  for  $d \geq 4$ ? ◊

We now remove the assumption that  $A$  is symmetric. Definition 12.16 makes sense for any  $A \in (\mathbb{R}^n)^{\otimes d}$ . That way, we lose the interpretations above, but we have a more general definition. Let us count the number of eigenpairs of a general tensor. The following formula is due to Cartwright and Sturmfels in [41].

**Theorem 12.18** Let  $A \in (\mathbb{R}^n)^{\otimes d}$  be general. The number of eigenvectors of  $A$ , up to scaling  $\mathbf{v} \mapsto t\mathbf{v}$ , is

$$\sum_{i=0}^{n-1} (d-1)^i.$$

**Proof** For  $d = 2$ , the sum gives  $n$ , which is the number of eigenvectors of a general matrix  $A \in \mathbb{R}^{n \times n}$ . Hence, we assume  $d > 2$  from now on. We use the same proof strategy as for Corollary 3.15. Consider the symmetric tensor  $A$  with  $F_A(\mathbf{x}) = x_1^d + x_2^d + \dots + x_n^d$ . The equations for its eigenpairs are

$$x_1^{d-1} - \xi^{d-2} x_1 = x_2^{d-1} - \xi^{d-2} x_2 = \dots = x_n^{d-1} - \xi^{d-2} x_n = 0,$$

where  $\lambda = \xi^{d-2}$  is the eigenvalue. We count that this system of homogeneous polynomial equations has precisely  $(d-1)^n$  regular solutions in  $\mathbb{P}^n$ . By Bézout's theorem, a general system of  $n$  polynomials with degrees  $(d-1, \dots, d-1)$  also has  $(d-1)^n$  regular zeros. The Parameter Continuation Theorem 3.18 then implies that the analogous equations for a general tensor  $A \in (\mathbb{R}^n)^{\otimes d}$  also have  $(d-1)^n$  regular zeros.

We remove the extraneous solution in  $\mathbb{P}^n$  that is given by  $\xi = 1$  and  $\mathbf{x} = 0$ . The remaining  $(d - 1)^n - 1$  solutions come in clusters of  $d - 2$ , where  $\mathbf{x}$  and  $\xi$  runs over the  $(d - 2)$ -th roots of unity. After taking the image of the projection  $(\xi, \mathbf{x}) \mapsto \mathbf{x}$ , the number of distinct points  $\mathbf{x}$  we obtain in  $\mathbb{P}^{n-1}$  is the quotient

$$\frac{(d - 1)^n - 1}{d - 2} = \sum_{i=0}^{n-1} (d - 1)^i.$$

We conclude that this equals the number of eigenvectors of a general tensor of format  $n \times n \times \cdots \times n$ .  $\square$

For matrices, every  $n$ -tuple of linearly independent vectors can be eigenvectors of a matrix. For tensors this is not so. Abo, Sturmfels and Seigal proved in [1] that a set of  $d$  points  $\mathbf{v}_1, \dots, \mathbf{v}_d$  in  $\mathbb{C}^2$  is the eigenconfiguration of a symmetric tensor in  $S^d(\mathbb{C}^2)$  if and only if  $d$  is odd, or  $d = 2k$  is even and the differential operator  $(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2})^k$  annihilates the binary form  $f_{\mathbf{v}_i}(x, y) := \prod_{i=1}^d (b_i x - a_i y)$ , where  $\mathbf{v}_i = (a_i, b_i)$ , for all  $i = 1, \dots, d$ . The extension of this result to ternary forms was studied by Beorchia, Galuppi and Venturello in [17]. This study of eigenschemes continues to be an area of active research.

In this section, we showed that the ED problem for Segre varieties and Veronese varieties leads directly to singular vectors and eigenvectors of tensors. By the ED duality in Theorem 2.23, it is equivalent to study the ED problem for the dual hypersurfaces, which are the hyperdeterminant and the discriminant respectively. This, in turn, takes us to the geometry of condition numbers, as discussed in Chapter 9. A natural generalization arises from the Segre-Veronese varieties, whose polar degrees we saw in Theorem 5.13.

## 12.3 Volumes of Rank-One Varieties

We now shift gears and we examine the metric geometry of the Segre variety and the Veronese variety. Both are now taken to be complex projective varieties. To this end, we pass to complex projective space  $\mathbb{P}^N$ , where  $N = n_1 \cdots n_d - 1$ , or  $N = n^d - 1$  if the  $n_i$  are all equal. Our two complex projective varieties are

$$\mathcal{S}_{\mathbb{P}} := \{\mathbf{v}_1 \otimes \cdots \otimes \mathbf{v}_d \in \mathbb{P}^N \mid \mathbf{v}_i \in \mathbb{P}^{n_i-1}, 1 \leq i \leq d\} \quad \text{and} \quad \mathcal{V}_{\mathbb{P}} := \{\mathbf{v}^{\otimes d} \in \mathbb{P}^N \mid \mathbf{v}_i \in \mathbb{P}^{n-1}\}. \quad (12.14)$$

Thus,  $\mathcal{S}_{\mathbb{P}}$  consists of rank-one tensors  $\mathbf{v}_1 \otimes \cdots \otimes \mathbf{v}_d$ , where the  $\mathbf{v}_i \in \mathbb{C}^{n_i}$  are complex vectors, taken up to scaling by  $\mathbb{C}^*$ . Similarly,  $\mathcal{V}_{\mathbb{P}}$  consists of all vectors  $\mathbf{v}^{\otimes d}$  for  $\mathbf{v} \in \mathbb{C}^n$ , again taken up to scaling by  $\mathbb{C}^*$ .

Our aim is to compute the volumes of  $\mathcal{S}_{\mathbb{P}}$  and  $\mathcal{V}_{\mathbb{P}}$ . The first crucial step is to explain what this means. In Section 6.3 and Chapter 14 we discuss volumes of real semialgebraic sets. But, here this situation is different. In this section, our objects are lower-dimensional submanifolds in the complex projective space  $\mathbb{P}^N$ . How does one measure volumes in  $\mathbb{P}^N$ ? How is the volume of a submanifold defined and computed?

Our point of departure is the Hermitian inner product on the tensor space  $\mathbb{C}^{n_1 \times \cdots \times n_d}$ . This is defined by

$$\langle A, B \rangle_{\mathbb{C}} := \sum_{i_1=1}^{n_1} \cdots \sum_{i_d=1}^{n_d} \overline{a_{i_1, \dots, i_d}} \cdot b_{i_1, \dots, i_d}$$

We view  $\mathbb{C}^{n_1 \times \cdots \times n_d}$  as a real Euclidean space of dimension  $2n_1 \cdots n_d$ , where the Euclidean inner product is

$$(A, B) \mapsto \operatorname{Re}(\langle A, B \rangle_{\mathbb{C}}).$$

The same definitions make sense in  $\mathbb{C}^n$ , and we define the unit sphere in the Euclidean space  $\mathbb{C}^n$  as

$$\mathbb{S}^{2n-1} := \{\mathbf{a} \in \mathbb{C}^n \mid \langle \mathbf{a}, \mathbf{a} \rangle_{\mathbb{C}} = 1\}.$$

This sphere is a real manifold of real dimension  $\dim_{\mathbb{R}} \mathbb{S}^{2n-1} = 2n - 1$ . The projection  $\pi : \mathbb{S}^{2n-1} \rightarrow \mathbb{P}^{n-1}$  sends a vector  $\mathbf{a} \in \mathbb{S}^{2n-1}$  to its image in complex projective space. The fiber of the quotient map  $\pi$  has real dimension one: it is a circle. The tangent space of the sphere  $\mathbb{S}^{2n-1}$  at a point  $\mathbf{a}$  is the real vector space

$$T_{\mathbf{a}} \mathbb{S}^{2n-1} = \{ \mathbf{t} \in \mathbb{C}^n \mid \operatorname{Re}(\langle \mathbf{a}, \mathbf{t} \rangle_{\mathbb{C}}) = 0 \} \simeq \mathbb{R}^{2n-1}.$$

For  $\mathbf{t} \in \mathbb{C}^n$ , we have  $\operatorname{Re}(\langle \mathbf{a}, \mathbf{t} \rangle_{\mathbb{C}}) = 0$  if and only if  $\langle \mathbf{a}, \mathbf{t} \rangle_{\mathbb{C}} = 0$  or  $\mathbf{t} = \lambda \sqrt{-1} \cdot \mathbf{a}$  for some  $\lambda \in \mathbb{R}$ . This implies the following representation of tangent space of projective space at the point  $\mathbf{x} = \pi(\mathbf{a})$ , where  $\mathbf{a} \in \mathbb{S}^{2n-1}$ :

$$T_{\mathbf{x}} \mathbb{P}^{n-1} = \{ \mathbf{t} \in \mathbb{C}^n \mid \langle \mathbf{a}, \mathbf{t} \rangle_{\mathbb{C}} = 0 \} \simeq \mathbb{R}^{2n-2}. \quad (12.15)$$

In conclusion, we have shown that the complex projective space  $\mathbb{P}^{n-1}$  is a Riemannian manifold with respect to the Euclidean structure  $\operatorname{Re}(\langle \mathbf{a}, \mathbf{b} \rangle_{\mathbb{C}})$ . This induces a notion of volume for subsets of  $\mathbb{P}^{n-1}$ .

Our metric structures on  $\mathbb{S}^{2n-1}$  and  $\mathbb{P}^{n-1}$  have the property that the projection  $\pi$  is a Riemannian submersion. This implies that the  $m$ -dimensional real volume of a measurable subset  $U \subset \mathbb{P}^{n-1}$  is

$$\operatorname{vol}_m(U) = \frac{1}{2\pi} \operatorname{vol}_{m+1}(\pi^{-1}(U)),$$

since the preimage  $\pi^{-1}(\mathbf{x})$  for  $\mathbf{x} \in \mathbb{P}^{n-1}$  is a circle. For instance, the volume of projective space itself is

$$\operatorname{vol}_{2(n-1)}(\mathbb{P}^{n-1}) := \frac{1}{2\pi} \operatorname{vol}_{2n-1}(\mathbb{S}^{2n-1}) = \frac{\pi^{n-1}}{(n-1)!}. \quad (12.16)$$

In the following, we sometimes omit the subscript from  $\operatorname{vol}$  when the dimension is clear from the context.

**Proposition 12.19** *Let  $m := \dim_{\mathbb{R}} \mathcal{S}_{\mathbb{P}}$ . The  $m$ -dimensional volume of the Segre variety  $\mathcal{S}_{\mathbb{P}}$  in (12.14) is*

$$\operatorname{vol}(\mathcal{S}_{\mathbb{P}}) = \operatorname{vol}(\mathbb{P}^{n_1-1}) \cdots \operatorname{vol}(\mathbb{P}^{n_d-1}).$$

**Proof** The Segre map (12.7) for projective space is the smooth embedding  $\psi_{\mathbb{P}} : \mathbb{P}^{n_1-1} \times \cdots \times \mathbb{P}^{n_d-1} \rightarrow \mathcal{S}_{\mathbb{P}}$ . Let  $(\mathbf{x}_1, \dots, \mathbf{x}_d) \in \mathbb{P}^{n_1-1} \times \cdots \times \mathbb{P}^{n_d-1}$  and  $\mathbf{a}_i \in \mathbb{S}^{2n_i-1}$  with  $\pi(\mathbf{a}_i) = \mathbf{x}_i$  for  $i = 1, \dots, d$ . We also fix elements  $\mathbf{t}_i \in \mathbb{C}^{n_i}$  with  $\langle \mathbf{a}_i, \mathbf{t}_i \rangle_{\mathbb{C}} = 0$  for all  $i$ . These specify vectors in the various tangent spaces (12.15).

The derivative of the Segre map  $\psi_{\mathbb{P}}$  at our point takes  $(\mathbf{t}_1, \dots, \mathbf{t}_d) \in T_{(\mathbf{x}_1, \dots, \mathbf{x}_d)}(\mathbb{P}^{n_1-1} \times \cdots \times \mathbb{P}^{n_d-1})$  to

$$\theta := \mathbf{t}_1 \otimes \mathbf{a}_2 \otimes \cdots \otimes \mathbf{a}_d + \mathbf{a}_1 \otimes \mathbf{t}_2 \otimes \cdots \otimes \mathbf{a}_d + \cdots + \mathbf{a}_1 \otimes \mathbf{a}_2 \otimes \cdots \otimes \mathbf{t}_d.$$

It follows from (12.3) that the terms in this sum are pairwise orthogonal. Therefore,

$$\|\theta\|^2 = \|\mathbf{t}_1\|^2 + \|\mathbf{t}_2\|^2 + \cdots + \|\mathbf{t}_d\|^2.$$

This shows that the derivative of  $\psi_{\mathbb{P}}$  preserves norms. This implies that  $\psi_{\mathbb{P}}$  is a volume-preserving embedding.  $\square$

**Proposition 12.20** *The  $2(n-1)$ -dimensional volume of the projective Veronese variety in (12.14) is*

$$\operatorname{vol}(\mathcal{V}_{\mathbb{P}}) = d^{n-1} \cdot \operatorname{vol}(\mathbb{P}^{n-1}).$$

**Proof** The proof is similar to that of Proposition 12.19. We denote the projective Veronese map by  $\nu_{\mathbb{P}} : \mathbb{P}^{n-1} \rightarrow \mathcal{V}_{\mathbb{P}}$ . Just like the Segre map,  $\nu_{\mathbb{P}}$  is a smooth embedding. Let  $\mathbf{x} \in \mathbb{P}^{n-1}$  be a point in projective space and fix a representative  $\mathbf{a} \in \mathbb{S}^{2n-1}$  for  $\mathbf{x}$ ; i.e.,  $\pi(\mathbf{a}) = \mathbf{x}$ . Let  $\mathbf{t} \in \mathbb{C}^n$  with  $\langle \mathbf{a}, \mathbf{t} \rangle = 0$ . The derivative of  $\nu_{\mathbb{P}}$  at  $\mathbf{a}$  maps  $\mathbf{t} \in T_{\mathbf{a}} \mathbb{P}^{n-1}$  to

$$\theta := \mathbf{t} \otimes \mathbf{a} \otimes \cdots \otimes \mathbf{a} + \mathbf{a} \otimes \mathbf{t} \otimes \cdots \otimes \mathbf{a} + \cdots + \mathbf{a} \otimes \mathbf{a} \otimes \cdots \otimes \mathbf{t}.$$

It follows from (12.3) that the terms in this sum are pairwise orthogonal, so  $\|\theta\|^2 = d\|\mathbf{t}\|^2$ . This shows that the derivative of  $v_{\mathbb{P}}$  scales norms by a factor of  $\sqrt{d}$ . This implies the desired formula

$$\text{vol}(\mathcal{V}_{\mathbb{P}}) = (\sqrt{d})^{\dim_{\mathbb{R}} \mathbb{P}^{n-1}} \cdot \text{vol}(\mathbb{P}^{n-1}) = d^{n-1} \cdot \text{vol}(\mathbb{P}^{n-1}),$$

which relates the volume of the projective Veronese variety to that of the underlying projective space.  $\square$

The volume of a complex projective variety is closely related to its degree. This is the content of Howard's *Kinematic Formula* [99], which concerns the average volume of intersections of submanifolds in homogeneous spaces. For complex projective space, we apply [99, Theorem 3.8 & Corollary 3.9] as follows. Let  $M \subset \mathbb{P}^N$  be a smooth manifold of complex dimension  $m$ . Here, the Kinematic Formula states

$$\mathbb{E}_U \#(M \cap U \cdot (\mathbb{P}^{N-m} \times \{0\}^m)) = \frac{\text{vol}_{2m}(M)}{\text{vol}_{2m}(\mathbb{P}^m)},$$

where the expectation refers to the probability Haar measure on the unitary group  $U(N+1)$ . If  $X \subset \mathbb{P}^N$  is a smooth variety of complex dimension  $m$ , then the number of intersection points  $\#(X \cap g \cdot (\mathbb{P}^{N-m} \times \{0\}^m))$  equals the degree of  $X$  for almost all  $g \in U(N+1)$ . Thus, we take the expected value of a constant function:

$$\deg(X) = \frac{\text{vol}_{2m}(X)}{\text{vol}_{2m}(\mathbb{P}^m)}, \quad \text{where } m = \dim_{\mathbb{C}}(X). \quad (12.17)$$

Using Propositions 12.19 and 12.20, we now derive the following result from our volume computation:

**Corollary 12.21** *The degrees of the Segre variety and the Veronese variety are equal to*

1.  $\deg(\mathcal{S}_{\mathbb{P}}) = \frac{(n_1 + \cdots + n_d - d)!}{(n_1 - 1)! \cdots (n_d - 1)!},$
2.  $\deg(\mathcal{V}_{\mathbb{P}}) = d^{n-1}.$

**Proof** The second formula follows directly from Proposition 12.20 and (12.17). For the first formula, we recall from (12.16) the volume of projective space:  $\text{vol}_{2(n-1)}(\mathbb{P}^{n-1}) = \frac{\pi^{n-1}}{(n-1)!}$ . We set

$$m := n_1 + \cdots + n_d - d = \dim_{\mathbb{C}} \mathcal{S}_{\mathbb{P}}.$$

Using Proposition 12.19 and (12.17) we then have

$$\deg(\mathcal{S}_{\mathbb{P}}) = \frac{\text{vol}(\mathbb{P}^{n_1-1}) \cdots \text{vol}(\mathbb{P}^{n_d-1})}{\text{vol}(\mathbb{P}^m)} = \frac{\pi^{\sum_{i=1}^d (n_i-1)}}{\pi^m} \frac{m!}{(n_1-1)! \cdots (n_d-1)!} = \frac{(n_1 + \cdots + n_d - d)!}{(n_1 - 1)! \cdots (n_d - 1)!}.$$

**Remark 12.22** By (12.2), a linear equation of  $\deg(\mathcal{V}_{\mathbb{P}})$  is homogeneous equation of degree  $d$  on  $\mathbb{P}^{n-1}$ . Thus,  $\deg(\mathcal{V}_{\mathbb{P}}) = d^{n-1}$  means that  $n-1$  general homogeneous polynomials of degree  $d$  have  $d^{n-1}$  zeros.

**Remark 12.23** Howard's Kinematic Formula [99] also provides the following result for real projective space  $\mathbb{P}_{\mathbb{R}}^N$ . Let  $M \subset \mathbb{P}_{\mathbb{R}}^N$  be a real submanifold of real dimension  $m$ . Then,  $\mathbb{E}_U \#(M \cap U \cdot (\mathbb{P}_{\mathbb{R}}^{N-m} \times \{0\}^m)) = \frac{\text{vol}_m(M)}{\text{vol}_m(\mathbb{P}_{\mathbb{R}}^m)}$ , where the expectation refers to the probability measure on the real orthogonal group  $O(n+1)$ . Thus, the volume of real projective varieties can be interpreted as an "average degree". We can use the same proof strategies as above to show that the projective volume of the real Segre variety  $\mathcal{S}$  is equal to  $\text{vol}(\mathbb{P}_{\mathbb{R}}^{n_1-1}) \cdots \text{vol}(\mathbb{P}_{\mathbb{R}}^{n_d-1})$  and the projective volume of the real Veronese  $\mathcal{V}$  is  $\sqrt{d^{n-1}} \cdot \text{vol}(\mathbb{P}_{\mathbb{R}}^{n-1})$ . The latter result was first derived by Edelman and Kostlan in their seminal paper [69]. We can use this now to find the average number of real zeros of a system of homogeneous polynomials of degree  $d$  in  $n$  variables.



## **Chapter 13**

## **Computer Vision**

The research field of computer vision studies how computers can gain understanding from images and videos, similarly to human cognitive abilities. One of the classical challenges is to reconstruct a 3D object from images taken by several unknown cameras. While the resulting questions in multiview geometry [85] have a long history in computer vision, the recent years have seen a confluence with ideas and algorithms from algebraic geometry. This led to the development of a new subject area called *Algebraic Vision* [109]. In this chapter we present an introduction to this area from the perspective of metric algebraic geometry.

### 13.1 Multiview Varieties

A pinhole camera consists of a point  $\mathbf{c}$ , called *camera center*, and an image plane  $H$  in 3-space. Taking an image of a 3D point  $\mathbf{x}$  is modeled by the following process. One draws a line between  $\mathbf{x}$  and the camera center  $\mathbf{c}$ . The intersection point of that line with  $H$  is the image point. This is illustrated in Figure 13.1.

An abstract algebraic model for a camera is a surjective linear projection  $C : \mathbb{P}^3 \dashrightarrow \mathbb{P}^2$ , given by an arbitrary full-rank  $3 \times 4$  matrix  $A$ , i.e.,  $C(\mathbf{x}) = A\mathbf{x}$ . Thus we work in projective geometry, with homogeneous coordinates. We consider 3D points in the projective space  $\mathbb{P}^3$ , and we identify the image plane  $H$  with the projective plane  $\mathbb{P}^2$ . In computer vision, one calls  $A$  a *projective camera* or *uncalibrated camera*.

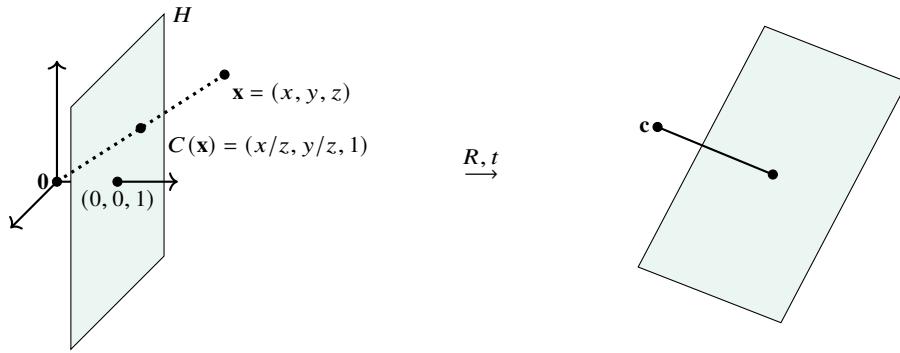


Fig. 13.1: Calibrated cameras: in standard position (on the left) and arbitrary (on the right).

In most real-life applications, one does not allow arbitrary matrices  $A$ , because one assumes several internal camera parameters, such as focal length, to be known. *Calibrated cameras* are those that can be obtained from rotation and translation from a camera in standard position; see Figure 13.1. Projectively, the standard camera is  $\mathbb{P}^3 \dashrightarrow \mathbb{P}^2$ ,  $[x_1 : x_2 : x_3 : x_4] \mapsto [x_1 : x_2 : x_3]$ , given by the  $3 \times 4$  matrix  $A = [I_3 | 0]$ . Thus, all calibrated cameras are given by matrices  $A = [R | t]$ , where  $R \in \text{SO}(3)$  and  $t \in \mathbb{R}^3$ . Between uncalibrated and calibrated cameras, there are also partially calibrated camera models; see [85, Chapter 6].

For 3D reconstruction from given images, one needs at least two cameras. The *joint camera map* is

$$\Phi : C_m \times X \dashrightarrow Y \tag{13.1}$$

that maps an  $m$ -tuple of cameras  $(C_1, \dots, C_m) \in C_m$  and an algebraic 3D object  $\mathbf{x} \in X$  to the  $m$  images  $(C_1(\mathbf{x}), \dots, C_m(\mathbf{x})) \in Y$ . Here  $X$  and  $Y$  are appropriate spaces of objects and their images respectively. For instance, the 3D object  $\mathbf{x}$  could be a single point (i.e.,  $X = \mathbb{P}^3$ ), a line (i.e.,  $X = \text{Gr}(1, \mathbb{P}^3)$ ), an arrangement of points and lines with prescribed incidences, a curve, or a surface. The image space  $Y$  parametrizes  $m$ -tuples of points, lines or other arrangements in  $\mathbb{P}^2$ . For instance, if  $X = \mathbb{P}^3$  then  $Y = (\mathbb{P}^2)^m$ .

Formally, the task of 3D reconstruction means to compute fibers under the joint camera map  $\Phi$ . Thus, there are unknown objects in 3-space and  $m$  unknown cameras. We get to see the planar projections of the objects. The task is to find the original objects and the cameras with which the images were taken.

**Example 13.1** For  $m = 2$  projective cameras observing  $k$  points, the joint camera map is

$$\begin{aligned} \Phi : (\mathbb{P}\mathbb{R}^{3 \times 4})^2 \times (\mathbb{P}_{\mathbb{R}}^3)^k &\dashrightarrow (\mathbb{P}_{\mathbb{R}}^2)^k \times (\mathbb{P}_{\mathbb{R}}^2)^k, \\ (A_1, A_2, \mathbf{x}_1, \dots, \mathbf{x}_k) &\mapsto (A_1 \mathbf{x}_1, \dots, A_1 \mathbf{x}_k, A_2 \mathbf{x}_1, \dots, A_2 \mathbf{x}_k). \end{aligned} \quad (13.2)$$

It is defined whenever none of the 3D points  $\mathbf{x}_i$  is the kernel of  $A_1$  nor  $A_2$ . The kernel of the matrix  $A_i$  is the *camera center* of  $C_i$ . Here,  $\Phi$  maps a space of dimension  $22 + 3k$  to a space of dimension  $4k$ . The general non-empty fiber has dimension at least 15, given by setting  $A_i \mapsto A_i g$  and  $\mathbf{x}_j \mapsto g^{-1} \mathbf{x}_j$  for some  $g \in \mathrm{PGL}(4)$ . For large  $k$ , the image of  $\Phi$  is a proper subvariety of its codomain. For small  $k$ , the map  $\Phi$  is dominant. The transition occurs for  $k = 7$ . This is the largest  $k$  such that  $\Phi$  is dominant. We will study this in Section 13.3.

For calibrated cameras, given by matrices of the form  $A_i = [R_i \mid t_i]$ , the joint camera map becomes

$$\Phi : (\mathrm{SO}(3) \times \mathbb{R}^3)^2 \times (\mathbb{P}_{\mathbb{R}}^3)^k \dashrightarrow (\mathbb{P}_{\mathbb{R}}^2)^k \times (\mathbb{P}_{\mathbb{R}}^2)^k. \quad (13.3)$$

Here, the domain has dimension  $12 + 3k$ , and the dominance transition occurs at  $k = 5$  (see Section 13.3).

We now change the problem, and we assume that the cameras are fixed and known. We are thus fixing a configuration  $C = (C_1, \dots, C_m) \in \mathcal{C}_m$ , and we consider the specialization  $\Phi_C : X \dashrightarrow Y$  of the joint camera map. The Zariski closure of the image of that map is the *multiview variety*  $\mathcal{M}_C$  of the cameras  $C$ .

In what follows, we focus on the multiview variety of a single point in 3-space. This means that  $X = \mathbb{P}^3$  and  $Y = (\mathbb{P}^2)^m$ . In this case, the multiview variety  $\mathcal{M}_C$  is the closure of the image of the map

$$\Phi_C : \mathbb{P}^3 \dashrightarrow (\mathbb{P}^2)^m, \quad \mathbf{x} \mapsto (A_1 \mathbf{x}, A_2 \mathbf{x}, \dots, A_m \mathbf{x}). \quad (13.4)$$

This map is well-defined at all points of  $\mathbb{P}^3$  except for the camera centers  $\ker(A_i)$ . If  $m \geq 2$  and not all camera centers coincide, then  $\mathcal{M}_C$  is a threefold. For an algebraic geometer,  $\mathcal{M}_C$  is the threefold obtained from  $\mathbb{P}^3$  by blowing up the camera centers. From a computational point of view, the first question is to find the implicit description of  $\mathcal{M}_C$  as a subvariety of  $(\mathbb{P}^2)^m$ . Writing  $(\mathbf{y}_1, \dots, \mathbf{y}_m)$  for the  $3m$  coordinates of  $(\mathbb{P}^2)^m$ , we seek multihomogeneous polynomials in the  $\mathbf{y}_i$  whose common zero set is equal to  $\mathcal{M}_C$ .

The desired implicit description is given by the  $3m \times (m+4)$  matrix

$$M_A := \begin{bmatrix} A_1 & \mathbf{y}_1 & 0 & \cdots & 0 \\ A_2 & 0 & \mathbf{y}_2 & \cdots & 0 \\ \vdots & & & \ddots & \\ A_m & 0 & 0 & \cdots & \mathbf{y}_m \end{bmatrix}. \quad (13.5)$$

Here  $A_i$  is the  $3 \times 4$  matrix that defines the camera  $C_i$ , and  $\mathbf{y}_i$  is an unknown column vector of length 3, serving as homogeneous coordinates for the  $i$ th image plane  $\mathbb{P}^2$ . The following result goes back to [92]. For recent ideal-theoretic versions of Proposition 13.2 see [109, Theorem 2] and the references therein.

**Proposition 13.2** *The multiview variety of a single point under  $m$  cameras with  $\geq 2$  distinct centers is*

$$\mathcal{M}_C = \{(\mathbf{y}_1, \dots, \mathbf{y}_m) \in (\mathbb{P}^2)^m \mid \mathrm{rk} M_A(\mathbf{y}_1, \dots, \mathbf{y}_m) < m+4\}.$$

**Proof (Idea)** A tuple  $(\mathbf{y}_1, \dots, \mathbf{y}_m) \in (\mathbb{P}^2)^m$  is in the image of (13.4) if and only if there exists  $\mathbf{x} \in \mathbb{P}^3$  and nonzero scalars  $\lambda_i$  such that  $A_i \mathbf{x} = \lambda_i \mathbf{y}_i$  for  $i = 1, \dots, m$ . The latter condition is equivalent to the vector

$(\mathbf{x}^\top, -\lambda_1, \dots, -\lambda_m)^\top$  being in the kernel of the matrix  $M_A$  in (13.5). Taking the closure, we obtain that the multiview variety  $\mathcal{M}_C$  consists of those tuples  $(\mathbf{y}_1, \dots, \mathbf{y}_m)$  such that  $M_A$  has non-zero kernel.  $\square$

**Example 13.3** For  $m = 2$  cameras, the matrix in (13.5) has format  $6 \times 6$ , and it equals

$$M_A = \begin{bmatrix} A_1 & \mathbf{y}_1 & 0 \\ A_2 & 0 & \mathbf{y}_2 \end{bmatrix}.$$

The multiview variety  $\mathcal{M}_C$  is the hypersurface in  $\mathbb{P}^2 \times \mathbb{P}^2$  defined by  $\det(M_A) = 0$ . This determinant is a bilinear form. Hence there exists a  $3 \times 3$  matrix  $F$ , whose entries depend on the matrices  $A_i$ , such that

$$\det(M_A) = \mathbf{y}_2^\top F \mathbf{y}_1. \quad (13.6)$$

The matrix  $F$  is called the *fundamental matrix*. A computation verifies the following explicit formula:

$$F = \begin{bmatrix} [2356] & -[1356] & [1256] \\ -[2346] & [1346] & -[1246] \\ [2345] & -[1345] & [1245] \end{bmatrix}. \quad (13.7)$$

Here,  $[ijkl]$  is the determinant of the  $4 \times 4$  submatrix with row indices  $i, j, k, l$  of the  $6 \times 4$  matrix  $\begin{bmatrix} A_1 \\ A_2 \end{bmatrix}$ . The matrix  $F$  has rank two. In fact, its right kernel is the image of the camera center of  $C_2$  under the camera  $C_1$ . Similarly, the left kernel of  $F$  is the image of the camera center of  $C_1$  under the camera  $C_2$ . A geometric interpretation is shown in Figure 13.2. The two cameras are projections onto the two planes.

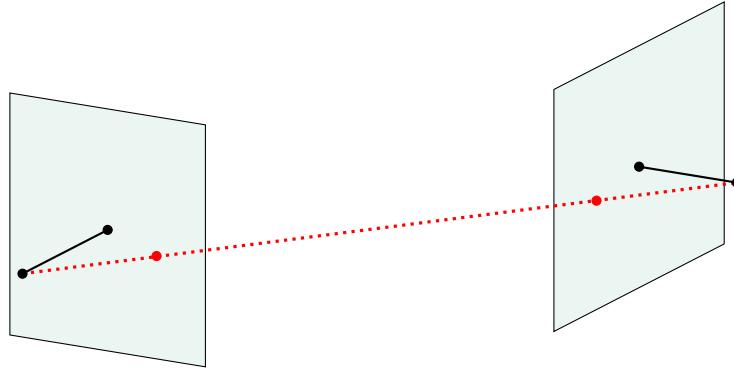


Fig. 13.2: The image of the center of one camera under the other camera (red points) is the kernel of the fundamental matrix.

In computer vision, the term *triangulation* is used quite differently from the usage in geometry and topology. Here, triangulation refers to the task of 3D reconstruction, assuming the configuration  $C$  of cameras is known. This task amounts to computing fibers under the specialized joint camera map  $\Phi_C$ .

Recall that, for  $m \geq 2$  cameras, the generic fiber under  $\Phi_C$  in (13.4) is empty. The measurements  $\mathbf{y} := (\mathbf{y}_1, \dots, \mathbf{y}_m)$  on the images are typically noisy. They do not lie on the multiview variety  $\mathcal{M}_C$ , but hopefully close it. To triangulate a corresponding space point  $\mathbf{x} \in \mathbb{P}^3$ , one seeks to compute the point  $\tilde{\mathbf{y}}$  on  $\mathcal{M}_C$  that is closest to  $\mathbf{y}$ , and then identify its fiber  $\Phi_C^{-1}(\tilde{\mathbf{y}})$ . Hence, the algebraic complexity of triangulating a single point from  $m$  images is determined by the ED degree of the multiview variety  $\mathcal{M}_C$ .

In real-life applications, the image measurements  $(\mathbf{y}_1, \dots, \mathbf{y}_m)$  come from an affine chart  $(\mathbb{R}^2)^m$  inside  $(\mathbb{P}^2)^m$ . Thus, we are working with the *affine multiview variety*  $\mathcal{M}_C^\circ := \mathcal{M}_C \cap (\mathbb{R}^2)^m$  in that affine chart. Triangulation thus means solving the ED problem (2.1) for the affine multiview variety  $\mathcal{M}_C^\circ$  in  $\mathbb{R}^{2m}$ .

The algebraic complexity of that problem is the ED degree of the affine variety  $\mathcal{M}_C^\circ$ . The interest in that ED degree originated in the computer vision community, namely in [84, 167]. Indeed, computer vision experts computed the ED degree of  $\mathcal{M}_C^\circ$  for  $m \leq 7$  cameras. Using Gröbner bases, one finds

$$\text{EDdegree}(\mathcal{M}_C^\circ) = 6, 47, 148, 336, 638, 1081 \quad \text{for } m = 2, 3, 4, 5, 6, 7. \quad (13.8)$$

These values were computed in the article [167], which is titled *How hard is 3-view triangulation really?* That computation was the original motivation for the development of the general notion of ED degrees of algebraic varieties. It started the five-author collaboration on the article [62]. Based on the numbers (13.8), the following cubic polynomial was conjectured in [62, Example 3.3] for the ED degree of the multiview variety  $\mathcal{M}_C^\circ$  for  $m$  cameras. That conjecture was proven by Maxim, Rodriguez and Wang in [130]:

**Theorem 13.4** *The ED degree of the affine multiview variety  $\mathcal{M}_C^\circ$  for  $m \geq 2$  cameras in general position is*

$$\frac{9}{2}m^3 - \frac{21}{2}m^2 + 8m - 4.$$

We close this section by explaining the Gröbner basis computation that solves the triangulation problem for small  $m$ . The  $m$  cameras are given by  $m$  matrices  $A_i = (a_{ijk})$  of format  $3 \times 4$ . In each image plane  $\mathbb{P}^2$ , we choose the affine chart by setting the last coordinate to 1. Our data are points  $(y_{i1}, y_{i2}, 1)$  in these planes for  $i = 1, 2, \dots, m$ . We similarly fix affine coordinates for the unknown point  $\mathbf{x} = (x_1, x_2, x_3, 1)$  in the space  $\mathbb{P}^3$ . Consider the squared Euclidean distance to a point  $\Phi_C(\mathbf{x})$  on  $\mathcal{M}_C^\circ$ , where  $\mathbf{x}$  is unknown:

$$\Delta(\mathbf{x}) = \sum_{i=1}^m \sum_{j=1}^2 \left( \frac{a_{ij1}x_1 + a_{ij2}x_2 + a_{ij3}x_3 + a_{ij4}}{a_{i31}x_1 + a_{i32}x_2 + a_{i33}x_3 + a_{i34}} - y_{ij} \right)^2.$$

The optimization problem (2.1) asks us to compute the point  $\mathbf{x}$  in  $\mathbb{R}^3$  that minimizes the distance  $\Delta(\mathbf{x})$ .

To model this problem algebraically, we consider the common denominator of our objective function:

$$\text{denom}(\mathbf{x}) := \prod_{i=1}^m (a_{i31}x_1 + a_{i32}x_2 + a_{i33}x_3 + a_{i34}).$$

We introduce a new variable  $z$  which satisfies  $z \cdot \text{denom}(\mathbf{x}) = 1$ . This ensures that all denominators are nonzero. Let  $I$  denote the ideal in the polynomial ring  $\mathbb{R}[x_1, x_2, x_3, z]$  that is generated by  $z \cdot \text{denom}(\mathbf{x}) - 1$  and the numerators of the partial derivatives  $\partial\Delta(\mathbf{x})/\partial x_k$  for  $k = 1, 2, 3$ . Thus  $I$  represents a system of four polynomial equations in four unknowns. The complex variety  $V(I)$  is finite, and its points are precisely the complex critical points of the objective function  $\Delta(\mathbf{x})$  in the open set given by  $\text{denom}(\mathbf{x}) \neq 0$ . We compute a Gröbner bases of  $I$  and we consider the set  $\mathcal{B}$  of standard monomials. By Proposition 3.7, we have  $\#\mathcal{B} = \#V(I)$ . That number equals  $6, 47, 148, \dots$  for  $m = 2, 3, 4, \dots$ , as promised in Theorem 13.4.

**Example 13.5** We consider the dinosaur data set from the Visual Geometry Group in the Department of Engineering Science at the University of Oxford. This data set contains 36 cameras as  $3 \times 4$  matrices and a list of pictures taken from 4983 3D points by these 36 cameras. The data is shown in Figure 13.3. We solve triangulation for  $m = 2$  randomly chosen cameras as described above, solving the resulting polynomial system with `HomotopyContinuation.jl` [29]. The result can be seen in Figure 13.4.

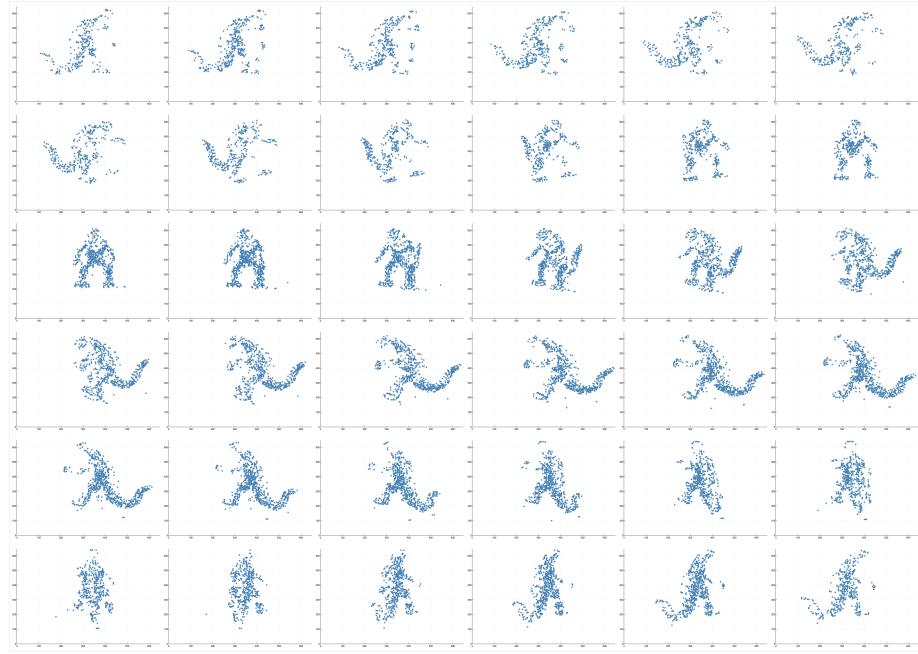


Fig. 13.3: The picture shows the 36 viewpoints on the 4983 3D points from the dinosaur data set in Example 13.5.

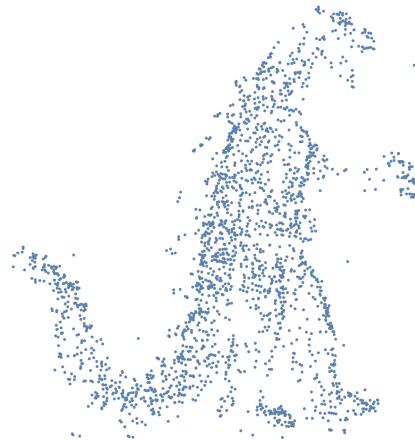


Fig. 13.4: The picture shows 4983 3D points that have been computed from the dinosaur data set in Example 13.5.

## 13.2 Grassmann Tensors

The fundamental matrix  $F$  from Example 13.3 is a projectively invariant representation of two cameras; see Proposition 13.12. Note that  $F$  has seven degrees of freedom, since  $F$  has rank two and is defined up to scale. This number makes sense because each matrix  $A_i$  has 11 degrees of freedom, up to scale, and we are taking the quotient modulo the 15-dimensional group  $\mathrm{PGL}(4)$ . Thus the dimension of the quotient space is  $2 \cdot 11 - 15 = 7$ .

In this section, we introduce a class of tensors which generalize the fundamental matrix. These give projectively invariant representations for  $m \geq 2$  cameras. We now allow higher-dimensional projections

$$C_i : \mathbb{P}^N \dashrightarrow \mathbb{P}^{n_i}. \quad (13.9)$$

Such projections are used to model basic dynamics [101, 183]. The *camera center* of the camera  $C_i$  is its base locus in  $\mathbb{P}^N$ . This is represented by the kernel of the corresponding matrix  $A_i$  with  $N + 1$  columns. The approach we shall describe was introduced in computer vision by Hartley and Schaffalitzky [83].

As before, the multiview variety is the closure in  $\mathbb{P}^{n_1} \times \cdots \times \mathbb{P}^{n_m}$  of the image of  $\mathbb{P}^N$  under the  $C_i$ . We are interested in tuples of linear spaces  $L_i \subset \mathbb{P}^{n_i}$  whose product meets the multiview variety. To this end, we fix the cameras  $C_i$  in (13.9), and we fix  $c_i = \text{codim}(L_i)$  with  $1 \leq c_i \leq n_i$ . Let  $\Gamma_{C,c}$  be the closure of

$$\begin{aligned} \{ (L_1, \dots, L_m) \in \text{Gr}(n_1 - c_1, \mathbb{P}^{n_1}) \times \dots \times \text{Gr}(n_m - c_m, \mathbb{P}^{n_m}) \mid \\ \exists \mathbf{x} \in \mathbb{P}^N : C_1(\mathbf{x}) \in L_1, \dots, C_m(\mathbf{x}) \in L_m \}. \end{aligned} \quad (13.10)$$

Thus,  $\Gamma_{C,c}$  is a subvariety in a product of Grassmannians. It fills its ambient space when the  $c_i$  are small. It has high codimension when the  $c_i$  are large. The following theorem concerns the sweet spot in the middle.

**Theorem 13.6** *For generic cameras  $C_i$ , the variety  $\Gamma_{C,c}$  is a hypersurface if and only if  $c_1 + \dots + c_m = N + 1$ . In that case, its defining equation is multilinear in the Plücker coordinates of the  $\text{Gr}(n_i - c_i, \mathbb{P}^{n_i})$ .*

Multilinear forms are represented by tensors, just like bilinear forms are represented by matrices. The equation described in Theorem 13.6 is a multilinear form in  $m$  sets of unknowns, namely the Plücker coordinates of the Grassmannians  $\text{Gr}(n_i - c_i, \mathbb{P}^{n_i})$  for  $i = 1, \dots, m$ . The *Grassmann tensor* of the camera configuration  $C$  is the order  $m$  tensor which represents the multilinear equation of the hypersurface  $\Gamma_{C,c}$ .

Before we come to the proof of Theorem 13.6, we present a census of all Grassmann tensors in the case of primary interest in computer vision, namely  $N = 3$  and  $n_i = 2$  for all  $i = 1, \dots, m$ . According to Theorem 13.6, a Grassmann tensor exists whenever  $c_1 + \dots + c_m = 4$  and  $1 \leq c_i \leq 2$  for  $i = 1, \dots, m$ .

**Example 13.7** ( $m = 2, c_1 = c_2 = 2$ ) Here the Grassmann tensor is the fundamental matrix  $F$ . The ambient space in (13.10) is  $(\text{Gr}(2 - 2, \mathbb{P}^2))^2 = (\mathbb{P}^2)^2$  with coordinates  $(\mathbf{y}_1, \mathbf{y}_2)$ , one for each of the two image points. The hypersurface  $\Gamma_{(C_1, C_2), (2, 2)}$  is the multiview variety  $\mathcal{M}_{C_1, C_2}$  in Example 13.3. The bilinear form in (13.6) is represented by the  $3 \times 3$  matrix  $F$ . The formula (13.7) can be obtained by computing the determinant of the  $6 \times 6$  matrix  $M_A$  using the Laplace expansion with respect to the last two columns.

**Example 13.8** ( $m = 3, c_1 = 2, c_2 = c_3 = 1$ ) This Grassmann tensor has format  $3 \times 3 \times 3$ . It is known as the *trifocal tensor* of three cameras. The ambient space in (13.10) is  $\text{Gr}(2 - 2, \mathbb{P}^2) \times (\text{Gr}(2 - 1, \mathbb{P}^2))^2 \cong (\mathbb{P}^2)^3$ . The first plane  $\mathbb{P}^2$  has coordinates  $\mathbf{y}_1$ . The other factors represent lines in  $\mathbb{P}^2$ , each spanned by two points  $\mathbf{y}_{i1}, \mathbf{y}_{i2}$  for  $i = 2, 3$ . The Plücker coordinates of the line  $\mathbf{y}_{i1} \vee \mathbf{y}_{i2}$  are the  $2 \times 2$  minors of the  $3 \times 2$  matrix  $[\mathbf{y}_{i1} \ \mathbf{y}_{i2}]$ . The determinant of the following  $9 \times 9$  matrix is a trilinear form in  $\mathbf{y}_1, \mathbf{y}_{21} \vee \mathbf{y}_{22}$  and  $\mathbf{y}_{31} \vee \mathbf{y}_{32}$ :

$$\det \begin{bmatrix} A_1 & \mathbf{y}_1 & 0 & 0 & 0 & 0 \\ A_2 & 0 & \mathbf{y}_{21} & \mathbf{y}_{22} & 0 & 0 \\ A_3 & 0 & 0 & 0 & \mathbf{y}_{31} & \mathbf{y}_{32} \end{bmatrix}.$$

There are actually three trifocal tensors, one for each choice of  $(c_1, c_2, c_3) \in \{(2, 1, 1), (1, 2, 1), (1, 1, 2)\}$ . The variety of all trifocal tensors, for one such choice fixed, was studied by Aholt and Oeding in [3].

**Example 13.9** ( $m = 4, c_1 = c_2 = c_3 = c_4 = 1$ ) Here we get the *quadrifocal tensor* of format  $3 \times 3 \times 3 \times 3$ . It represents a multilinear form on four sets of lines  $\mathbf{y}_{i1} \vee \mathbf{y}_{i2}$ , namely the determinant of the  $12 \times 12$  matrix

$$\det \begin{bmatrix} A_1 & \mathbf{y}_{11} & \mathbf{y}_{12} & 0 & 0 & 0 & 0 & 0 & 0 \\ A_2 & 0 & 0 & \mathbf{y}_{21} & \mathbf{y}_{22} & 0 & 0 & 0 & 0 \\ A_3 & 0 & 0 & 0 & 0 & \mathbf{y}_{31} & \mathbf{y}_{32} & 0 & 0 \\ A_4 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{y}_{41} & \mathbf{y}_{42} \end{bmatrix}.$$

The variety of all quadrifocal tensors was studied by Oeding in [140].

**Proof (of Theorem 13.6)** This result appears in [83]. We give an independent proof. For general  $L_i \in \text{Gr}(n_i - c_i, \mathbb{P}^{n_i})$ , the back-projected space  $\tilde{L}_i := C_i^{-1}(L_i)$  also has codimension  $c_i$ , i.e.,  $\tilde{L}_i \in \text{Gr}(N - c_i, \mathbb{P}^N)$ . If  $\sum c_i \leq N$ , the subspaces  $\tilde{L}_1, \dots, \tilde{L}_m$  always intersect, meaning that  $\Gamma_{C,c}$  is equal to its ambient space.

Now suppose  $\sum c_i = N + 1$ . Due to the genericity of the cameras, the multiview variety  $\mathcal{M}_C$  of a single point (i.e. the Zariski closure of the image of  $\Phi_C : \mathbb{P}^N \dashrightarrow \mathbb{P}^{n_1} \times \cdots \times \mathbb{P}^{n_m}$ ) has dimension  $N$ . Hence, the product  $L_1 \times \cdots \times L_m$  of general subspaces  $L_i \subseteq \mathbb{P}^{n_i}$  of codimension  $c_i$  does not intersect the multiview variety  $\mathcal{M}_C$ . Therefore,  $\Gamma_{C,c}$  is a proper subvariety. We show that it is a hypersurface and simultaneously compute its multidegree by intersecting it with generic pencils in  $\text{Gr}(n_1 - c_1, \mathbb{P}^{n_1}) \times \cdots \times \text{Gr}(n_m - c_m, \mathbb{P}^{n_m})$ . It suffices to show that every general pencil meets  $\Gamma_{C,c}$  in exactly one point. Such a pencil is of the form  $L_1 \times \cdots \times \mathcal{L}_k \times \cdots \times L_m$ , where  $L_i$  (for  $i \neq k$ ) is a general point in the Grassmannian  $\text{Gr}(n_i - c_i, \mathbb{P}^{n_i})$ , and

$$\mathcal{L}_k = \{L \in \mathrm{Gr}(n_k - c_k, \mathbb{P}^{n_k}) \mid V_k \subset L \subset W_k\},$$

where  $V_k, W_k \subseteq \mathbb{P}^{n_k}$  are general subspaces of codimension  $c_k + 1$  and  $c_k - 1$ , respectively. The back-projected spaces  $\tilde{L}_1, \dots, \tilde{W}_k, \dots, \tilde{L}_m$  intersect in a single point  $\mathbf{x} \in \mathbb{P}^N$ . That point  $\mathbf{x}$  does not lie in the back-projected space  $\tilde{V}_k$ , because otherwise, for a general  $L \in \mathcal{L}_k$ , the point  $L_1 \times \dots \times L \times \dots \times L_m$  would be in  $\Gamma_{C,c}$ . This is a contradiction since the codimension of  $\Gamma_{C,c}$  is at least one. Thus, there is exactly one  $L \in \mathcal{L}_k$  (namely, the span of  $V_k$  with  $C_k(\mathbf{x})$ ) with the property that  $L_1 \times \dots \times L \times \dots \times L_m \in \Gamma_{C,c}$ . This shows that  $\Gamma_{C,c}$  is a hypersurface of multidegree  $(1, \dots, 1)$ . The multidegree of a hypersurface in a product of Grassmannians equals the multidegree of its equation in Plücker coordinates [76, Chapter 3, Proposition 2.1]. Similar arguments show that the codimension of  $\Gamma_{C,c}$  is  $\geq 2$  whenever  $\sum c_i > N + 1$ .  $\square$

By varying the Grassmann tensors over all camera configurations, we obtain a variety that parametrizes camera configurations modulo projective transformations; see Proposition 13.12. In the classical case  $N = 3$ , this quotient space has dimension  $11m - 15$ . This approach yields the variety of fundamental matrices for  $m = 2$ , the trifocal variety [3] for  $m = 3$ , and the quadrifocal variety [140] for  $m = 4$ . For  $m \geq 5$ , we can form quotient varieties by taking appropriate collections of fundamental matrices, trifocal tensors and quadrifocal tensors. A more systematic representation is furnished by the *Hilbert schemes* due to Aholt, Sturmfels and Thomas [4].

### 13.3 3D Reconstruction from Unknown Cameras

We now turn to the problem of 3D reconstruction from images taken by *unknown* cameras. This amounts to computing fibers under the joint camera map  $\Phi$  in (13.1). Typically, a nontrivial group  $G$  acts on the fibers of  $\Phi$  since global 3D transformations that act simultaneously on the cameras and the 3D scene do not change the resulting images. Most real-life applications assume that the cameras are calibrated, so we are naturally led to the rotation group  $\text{SO}(3)$ , which is an important player in metric algebraic geometry.

**Example 13.10** For a projective camera  $A \in \mathbb{P}\mathbb{R}^{3 \times 4}$  that observes a point  $\mathbf{x} \in \mathbb{P}^3$ , the projective linear group  $\text{PGL}(4)$  acts via  $g \mapsto (Ag^{-1}, g\mathbf{x})$  on cameras and points without changing the image  $Ag^{-1} \cdot g\mathbf{x} = A\mathbf{x}$ . Hence,  $\text{PGL}(4)$  acts on the fibers of the joint camera map in (13.2), where two projective cameras observe  $k$  points in  $\mathbb{P}^3$ . This means that 3D reconstruction is only possible up to a projective transformation.

The action of  $\mathrm{PGL}(4)$  does not map calibrated cameras to calibrated cameras. The largest subgroup  $G$  of  $\mathrm{GL}(4)$  that preserves the structure of calibrated camera matrices  $[R \mid t] \in \mathrm{SO}(3) \times \mathbb{R}^3$  equals

$$G = \{g \in \mathrm{GL}(4) \mid g = \begin{bmatrix} R & t \\ 0 & \lambda \end{bmatrix} \text{ for some } R \in \mathrm{SO}(3), t \in \mathbb{R}^3, \lambda \in \mathbb{R} \setminus \{0\}\}. \quad (13.11)$$

This is the scaled special Euclidean group of  $\mathbb{R}^3$ . Paraphrasing for computer vision, we conclude that 3D reconstruction with calibrated cameras is only possible up to a proper rigid motion and a nonzero scale.

The state-of-the-art 3D reconstruction algorithms work with *minimal problems*, where the joint camera map  $\Phi$  is dominant and its fibers are generically finite modulo the group  $G$ . For minimal problems, computing the fibers of  $\Phi$  means solving a *square* parametrized system of polynomial equations. Such 3D reconstruction problems use the minimal amount of data on the images while having finitely many solutions. The *algebraic degree* of a minimal problem is the number of complex solutions for generic data.

**Proposition 13.11** *Consider 3D reconstruction for  $m \geq 2$  unknown cameras observing  $k$  points. For projective cameras, there are two minimal problems, namely  $(m, k) = (2, 7)$  or  $(3, 6)$ , both of algebraic degree 3. For calibrated cameras, the only minimal problem is  $(m, k) = (2, 5)$ , with algebraic degree 20.*

**Proof** For projective cameras, the map  $\Phi$  takes the quotient space  $((\mathbb{PR}^{3 \times 4})^m \times (\mathbb{P}_{\mathbb{R}}^3)^k)/\mathrm{PGL}(4)$  to the space of images  $((\mathbb{P}^2)^k)^m$ . (A comment for experts: the quotient is meant in the sense of *geometric invariant theory*, but we here only need a birational model.) If the 3D reconstruction problem is minimal, then both spaces have the same dimension, which gives  $11m + 3k - 15 = 2km$ . This equation has precisely two integer solutions  $m \geq 2$  and  $k \geq 1$ , namely  $(m, k) = (2, 7), (3, 6)$ . Both yield minimal problems of algebraic degree three [82]. These degrees were found in the 19th century by Hesse [91] and Sturm [169].

For calibrated cameras, the domain of the joint camera map  $\Phi$  changes to  $((\mathrm{SO}(3) \times \mathbb{R}^3)^m \times (\mathbb{P}^3)^k)/G$  where  $G$  is the special Euclidean group (13.11). Now, the two spaces have the same dimension if and only if  $6m + 3k - 7 = 2km$ . The only relevant integer solution is  $(m, k) = (2, 5)$ . This yields a minimal problem of algebraic degree 20. This problem has the label  $5000_2$  in the census of minimal problems found in [66, Table 1]. It is the minimal problem most widely used in practical 3D reconstruction algorithms.  $\square$

Later we shall take a closer look at the polynomial systems arising from Proposition 13.11. First, however, we explain the practical usage of minimal problems for 3D reconstruction in computer vision. Imagine two calibrated cameras that observe 100 points. Then the joint camera map  $\Phi$  in (13.3) for  $k = 100$  is not dominant. Hence, the fiber under  $\Phi$  of two noisy images is empty. We would need to find a closest point on the image of  $\Phi$ , before we can compute any fiber. In other words, we must first solve an ED problem like that in Theorem 13.4. But this is now even harder because the cameras are unknown. In addition, in practical scenarios, it often happens that some of the given point pairs are outliers. Then, solving the ED problem yields incorrect solutions.

In practice, one simply avoids any such ED problem. Instead, one chooses five of the 100 given point pairs and solves the associated minimal problem in Proposition 13.11. This process gets repeated many times. The solutions of the many minimal problems get patched together via *random sample consensus* (RANSAC), until all 100 points and both cameras are reconstructed. For more details see [109].

Since minimal problems must be solved many times for RANSAC, it is crucial that their formulation as a square polynomial system is efficient. A common strategy for simplification is to first only reconstruct the cameras and afterwards recover the 3D scene via triangulation. Hence, instead of solving the full minimal problem at once, we start with a polynomial system whose only unknowns are the camera parameters.

To parametrize the  $m$  cameras modulo the group  $G$ , we use the Grassmann tensors from Section 13.2. We work in arbitrary dimensions as in (13.10), starting with  $G = \mathrm{PGL}(N+1)$ . We consider  $m$  surjective projections  $C_i : \mathbb{P}^N \dashrightarrow \mathbb{P}^{n_i}$  and integers  $c_1, \dots, c_m$  that satisfy  $1 \leq c_i \leq n_i$  and  $c_1 + \dots + c_m = N+1$ . The Grassmann tensor  $T_{C,c}$  exists, by Theorem 13.6. Namely,  $T_{C,c}$  is parametrized by the rational map

$$\gamma_c : \mathbb{P}\mathbb{R}^{(n_1+1) \times (N+1)} \times \cdots \times \mathbb{P}\mathbb{R}^{(n_m+1) \times (N+1)} \dashrightarrow \mathbb{P}\mathbb{R}^{\binom{n_1+1}{c_1} \times \cdots \times \binom{n_m+1}{c_m}}$$

that sends the  $m$  matrices for the cameras  $C_i$  to their Grassmann tensor, written in Plücker coordinates.

**Proposition 13.12** ([83]) *The group  $\mathrm{PGL}(N+1)$  acts on the fibers of  $\gamma_c$  by componentwise right-multiplication. The map  $\gamma_c$  becomes birational on the quotient of its domain modulo that group action.*

**Example 13.13** Consider the basic scenario in Example 13.7, with  $N = 3$  and  $m = n_1 = n_2 = c_1 = c_2 = 2$ . Here, the map  $\gamma_c$  sends two  $3 \times 4$  matrices  $A_1, A_2$  to the  $3 \times 3$  matrix  $F$  in (13.7). The Zariski closure of the image of  $\gamma_c$  is the variety  $\mathcal{F} = V(\det)$  defined by the  $3 \times 3$  determinant. This is a hypersurface in  $\mathbb{RP}^{3 \times 3} = \mathbb{P}^8$ . It is birational to pairs of projective cameras modulo  $\mathrm{PGL}(4)$ . This quotient construction reduces the number of camera parameters from 22 to 7. Any pair of image points  $(\mathbf{y}_1, \mathbf{y}_2) \in \mathbb{P}^2 \times \mathbb{P}^2$  imposes a linear condition on  $\mathcal{F}$ , namely  $\mathbf{y}_2^\top F \mathbf{y}_1 = 0$  as in (13.6). Therefore, reconstructing a fundamental matrix means intersecting the seven-dimensional variety  $\mathcal{F}$  with seven hyperplanes. We conclude that observing seven points with two projective cameras is a minimal problem of algebraic degree  $\deg(\mathcal{F}) = 3$ .

We now turn to calibrated cameras and to the scaled special Euclidean group  $G$  in (13.11). A real  $3 \times 3$  matrix  $E$  is called *essential* if it has rank two and the two non-zero singular values are equal. The Zariski closure of the set of essential matrices is the *essential variety*. This lives in  $\mathbb{RP}^{3 \times 3} = \mathbb{P}^8$  and is denoted  $\mathcal{E}$ .

**Theorem 13.14** *The map  $\gamma_c$  is two-to-one when restricted to calibrated cameras modulo  $G$ . Its image is the essential variety  $\mathcal{E}$ . This has dimension five, degree ten, and is defined by the ten cubic equations*

$$\det E = 0 \quad \text{and} \quad EE^\top E - \frac{1}{2} \mathrm{tr}(EE^\top)E = 0. \quad (13.12)$$

**Proof** This can be proved by direct computations. The set of essential matrices is a semialgebraic set of codimension 3. Demazure [56] showed that the equations (13.12) hold on that set. With Macaulay2 we can check that these ten homogeneous cubics generate a prime ideal of codimension 3. The fact that the map  $\gamma_c$  is two-to-one can be verified by a computation in the coordinates that are used in [66].  $\square$

The reconstruction of calibrated cameras is now analogous to Example 13.13. Any pair of image points  $(\mathbf{y}_1, \mathbf{y}_2) \in \mathbb{P}^2 \times \mathbb{P}^2$  imposes a linear condition  $\mathbf{y}_2^\top E \mathbf{y}_1 = 0$  on  $\mathcal{E}$ . Hence we intersect the five-dimensional variety  $\mathcal{E}$  with five hyperplanes. That intersection consists of ten complex matrices  $E$ . Each  $E$  comes from two pairs of calibrated cameras modulo  $G$ . We conclude that observing five points with two calibrated cameras is a minimal problem of algebraic degree  $20 = 2 \cdot 10$ , confirming Proposition 13.11.

To reiterate: what we described is the standard approach in state-of-the-art reconstruction algorithms. One first reconstructs an essential matrix by intersecting the variety  $\mathcal{E}$  with five hyperplanes, then recovers the calibrated cameras, and finally reconstructs the 3D scene using triangulation. It would be interesting to study this process through the lens of metric algebraic geometry. In particular, following Chapter 9, what is the condition number of intersecting  $\mathcal{E}$  with five hyperplanes  $\mathbf{y}_2^\top E \mathbf{y}_1 = 0$  given by five image point pairs  $(\mathbf{y}_1, \mathbf{y}_2)$ ? One may hope to utilize those condition numbers in homotopy continuation solvers.

More generally, we can ask the same question for intersecting the variety  $\mathcal{G}_c := \overline{\mathrm{im}(\gamma_c)}$  of Grassmann tensors with  $\dim \mathcal{G}_c$  hyperplanes given by  $(L_1, \dots, L_m)$  as in (13.10)? The condition number of intersecting a fixed projective variety with varying linear subspaces of complementary dimension was studied in [38]. However, that theory does not immediately apply to our problem, since  $\mathcal{E}$  is only intersected by *special* linear spaces, namely those that are intersections of five hyperplanes of the form  $\{E : \mathbf{y}_2^\top E \mathbf{y}_1 = 0\}$ .

## **Chapter 14**

### **Volumes**

In this chapter we discuss the problem of computing the volume of a subset  $X$  of  $\mathbb{R}^n$  that is full-dimensional and semialgebraic. Being *semialgebraic* means that  $X$  is described by a finite Boolean combination of polynomial inequalities. We say that  $X$  is *basic semialgebraic* if that description is a conjunction of polynomial inequalities. This means that our set admits a representation of the form

$$X = \{ \mathbf{x} \in \mathbb{R}^n : f_1(\mathbf{x}) \geq 0 \text{ and } f_2(\mathbf{x}) \geq 0 \text{ and } \dots \text{ and } f_k(\mathbf{x}) \geq 0 \},$$

where  $f_1, f_2, \dots, f_k$  are polynomials in  $n$  unknowns with real coefficients. Our task is to compute the volume of  $X$ , as reliably and accurately as possible, when the input consists of the polynomials  $f_1, f_2, \dots, f_k$ .

## 14.1 Calculus and Beyond

The simplest scenario arises when  $k = 1$ , so our semialgebraic set  $X$  is the domain of nonnegativity of one polynomial  $f(\mathbf{x}) = f(x_1, \dots, x_n)$  with real coefficients. We wish to evaluate the integral

$$\text{Vol}(X) = \int_X 1 \cdot d\mathbf{x}, \quad (14.1)$$

where  $d\mathbf{x}$  denotes Lebesgue measure on  $\mathbb{R}^n$ . Of course, it makes perfect sense to also consider integrals  $\int_X g(\mathbf{x})d\mathbf{x}$ , where  $g(\mathbf{x})$  is some polynomial function. The value of such an integral is a real number which is called a *period* [118]. Our integrals are special cases of *period integrals*.

We begin with an instance where the volume can be computed explicitly using calculus.

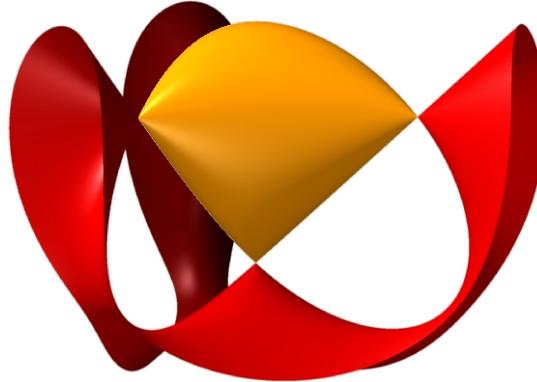


Fig. 14.1: The yellow convex body is the ellipote. It is bounded by Cayley's cubic surface.

**Example 14.1 (Ellipote)** Consider the set  $X$  of all points  $(x, y, z)$  in  $\mathbb{R}^3$  such that the matrix

$$M = \begin{pmatrix} 1 & x & y \\ x & 1 & z \\ y & z & 1 \end{pmatrix}$$

is positive semidefinite. The set  $X$  is convex and semialgebraic: it consists of all points  $(x, y, z)$  in the cube  $[-1, 1]^3$  such that  $\det(M) = 2xyz - x^2 - y^2 - z^2 + 1$  is nonnegative. Figure 14.1 appears in [135, Figure 1.1] and it serves as the logo of the Nonlinear Algebra group at the Max-Planck Institute for Mathematics

in the Sciences in Leipzig. It illustrates several applications of algebraic geometry. In statistics, the convex set  $X$  is the set of all correlation matrices. In optimization, it is the feasible region of a semidefinite programming problem [135, Chapter 12], and it is known as the *elliptope*.

We now compute the volume of the elliptope. We begin by rewriting of its boundary surface. Solving the equation  $\det(M) = 0$  for  $z$  with the quadratic formula, we obtain

$$z = xy \pm \sqrt{x^2y^2 - x^2 - y^2 + 1} = xy \pm \sqrt{(1-x^2)(1-y^2)} \quad \text{for } (x, y) \in [-1, 1]^2,$$

The plus sign gives the upper yellow surface and the minus sign gives the lower yellow surface. The volume of the elliptope  $X$  is obtained by integrating the difference between the upper function and the lower function over the square. Hence the desired volume equals

$$\text{vol}(X) = \int_{-1}^1 \int_{-1}^1 2 \sqrt{(1-x^2)(1-y^2)} dx dy = 2 \left[ \int_{-1}^1 \sqrt{1-t^2} dt \right]^2.$$

The univariate integral on the right gives the area of a semicircle with radius 1. We know from trigonometry that this area equals  $\pi/2$ , where  $\pi = 3.14159265\dots$ . We conclude that

$$\text{vol}(X) = \pi^2/2 = 4.934802202\dots$$

Thus our elliptope covers about 61.7 % of the volume of the cube  $[-1, 1]^3$  that surrounds it.

The number  $\pi^2/2$  we found is an example of a period. It is generally much more difficult to accurately evaluate such integrals. In fact, this challenge has played an important role in the history of mathematics. Consider the problem of computing the arc length of an ellipse. This requires us to integrate the reciprocal square root of cubic polynomial  $f(t)$ . Such integrals are called *elliptic integrals*, and they represent periods of elliptic curves. Furthermore, in an 1841 paper, Abel introduced *abelian integrals*, where  $g(t)$  is an algebraic function in one variable  $t$ . How to evaluate such an integral? This question leads us to Riemann surface and then to their Jacobians. And, voilà, we arrived at the theory of *abelian varieties*.

This chapter presents two current paradigms for accurately computing integrals like (14.1). The first method rests on the theory of  $D$ -modules, that is, on the algebraic study of linear differential equations with polynomial coefficients. Our volume is found as a special value of a parametric volume function that is encoded by means of its *Picard-Fuchs differential equation*. This method, which tends to appeal to algebraic geometers, was introduced by Lairez, Mezzarobba and Safey El Din in [119].

The second approach is due to Lasserre and his collaborators [90, 175, 176]. On first glance it might appeal more to analysts and optimizers, but there is also plenty of deep algebraic structure under the hood. The idea is to consider all moments  $m_{\mathbf{a}} = \int_X \mathbf{x}^{\mathbf{a}} d\mathbf{x}$  of our semialgebraic set  $X$ , where again  $d\mathbf{x}$  is Lebesgue measure, and to use relations among these moments to infer an accurate approximation of  $m_0 = \text{vol}(X)$ . That numerical inference rests on semidefinite programming [135, Chapter 12].

## 14.2 D-Modules

In calculus, we learn about definite integrals in order to determine the area under a graph. Likewise, in multivariable calculus, we examine the volume enclosed by a surface. We are here interested in areas and volumes of semi-algebraic sets. When these sets depend on one or more parameters, their volumes are holonomic functions of the parameters. We explain what this means and how it can be used for accurate evaluation of volume functions. We present the method of [119], following the exposition given in [164].

Suppose that  $M$  is a  $D$ -module. The letter  $D$  denotes the *Weyl algebra* (cf. [160, 164]), written here as

$$D = \mathbb{C}\langle x_1, \dots, x_n, \partial_1, \dots, \partial_n \rangle.$$

In applications,  $M$  is usually a space of infinitely differentiable functions on a subset of  $\mathbb{R}^n$  or  $\mathbb{C}^n$ . Such  $D$ -modules are torsion-free. For a function  $f \in M$ , its *annihilator* is the  $D$ -ideal

$$\text{Ann}_D(f) := \{P \in D \mid P \bullet f = 0\}.$$

In general, it is a non-trivial task to compute this annihilating ideal. But, in some cases, computer algebra systems can help us to compute holonomic annihilating ideals. For rational functions  $r \in \mathbb{Q}(x_1, \dots, x_n)$  this can be done in **Macaulay2** with a built-in command as follows:

```
needsPackage "Dmodules"; D = QQ[x1,x2,d1,d2, WeylAlgebra => {x1=>d1,x2=>d2}];  
rnum = x1; rden = x2; I = RatAnn(rnum,rden)
```

This code fragment shows that  $r = x_1/x_2$  has  $\text{Ann}_D(r) = D\{\partial_1^2, x_1\partial_1 - 1, x_2\partial_1\partial_2 + \partial_1\}$ .

Suppose now that  $f(x_1, \dots, x_n)$  is an algebraic function. This means that  $f$  satisfies some polynomial equation  $F(f, x_1, \dots, x_n) = 0$ . Using the polynomial  $F$  as its input, the **Mathematica** package **HolonomicFunctions** can compute a holonomic representation of  $f$ . The output is a linear differential operator of lowest degree annihilating  $f$ . See Example 14.3.

Let  $M$  be a  $D$ -module and  $f \in M$ . We say that  $f$  is *holonomic* if, for each  $i \in \{1, \dots, n\}$ , there is an operator  $P_i \in \mathbb{C}[x_1, \dots, x_n]\langle\partial_i\rangle \setminus \{0\}$  that annihilates  $f$ . If this holds then we say that  $\text{Ann}_D(f)$  is a *holonomic  $D$ -ideal*. Suppose this is the case, and fix a general point  $x_0 \in \mathbb{C}^n$ . Let  $m_1, \dots, m_n$  denote the orders of the differential operators  $P_1, \dots, P_n$  in the definition of holonomic. Thus,  $P_k$  is an operator in  $\partial_k$  of order  $m_k$  whose coefficients are polynomials in  $x_1, \dots, x_n$ .

We fix initial conditions for  $f$  by specifying the following  $m_1 m_2 \cdots m_n$  numerical values:

$$(\partial_1^{i_1} \cdots \partial_n^{i_n} \bullet f)|_{x=x_0} \quad \text{where } 0 \leq i_k < m_k \text{ for } k = 1, \dots, n. \quad (14.2)$$

Then the operators  $P_1, \dots, P_n$  together with the initial conditions (14.2) specify the function  $f$ .

Many interesting functions are holonomic. To begin with, every rational function  $r$  in  $\mathbf{x} = (x_1, \dots, x_n)$  is holonomic, because  $r$  is annihilated by  $r(\mathbf{x})\partial_i - \partial r / \partial x_i$  for  $i = 1, 2, \dots, n$ . By clearing denominators in this operator, we obtain a non-zero  $P_i \in \mathbb{C}[\mathbf{x}]\langle\partial_i\rangle$  with  $m_i = 1$  that annihilates  $r$ . See the **Macaulay2** example above. These operators, together with fixing the value  $r(\mathbf{x}_0)$  at a general point  $\mathbf{x}_0 \in \mathbb{C}^n$ , constitute a canonical holonomic representation of the rational function  $r$ .

Holonomic functions in one variable are solutions to ordinary linear differential equations with rational function coefficients. Examples include algebraic functions, some elementary trigonometric functions, hypergeometric functions, Bessel functions, period integrals, and many more. But, not every nice function is holonomic. A necessary condition for a meromorphic function  $f(x)$  to be holonomic is that it has only finitely many poles in  $\mathbb{C}$ . For a concrete example, we start with the holonomic function  $\sin(x)$ . This is annihilated by the operator  $\partial^2 + 1$ . Its reciprocal  $f(x) = \frac{1}{\sin(x)}$  has infinitely many poles, so is not holonomic. Hence the class of holonomic functions is not closed under division. It is also not closed under composition of functions, since both  $\frac{1}{x}$  and  $\sin(x)$  are holonomic. We record the following fact:

**Proposition 14.2** *Let  $f(\mathbf{x})$  be holonomic and  $g(\mathbf{x})$  algebraic. Then  $f(g(\mathbf{x}))$  is holonomic.*

For the proof see [164, Proposition 2.3]. The term ‘‘holonomic function’’ is due to Zeilberger [184]. Koutschan [117] developed practical algorithms for manipulating holonomic functions. These are implemented in his **Mathematica** package **HolonomicFunctions**, as seen below.

**Example 14.3** Every algebraic function  $f(\mathbf{x})$  in  $n$  variables is holonomic. Let  $n = 2$  and consider the function  $y = f(x)$  that is defined by  $y^4 + x^4 + \frac{xy}{100} - 1 = 0$ . Its annihilator in  $D$  can be computed as follows:

```
<< RISC`HolonomicFunctions`
q = y^4 + x^4 + x*y/100 - 1
ann = Annihilator[Root[q, y, 1], Der[x]]
```

This Mathematica code determines an operator  $P$  of lowest order in  $\text{ann}_D(f)$ . We find

$$\begin{aligned} P = & (2x^4 + 1)^2 (25600000000x^{12} - 76800000000x^8 + 76799999973x^4 - 25600000000) \partial^3 \\ & + 6x^3(2x^4 + 1)(51200000000x^{12} + 76800000000x^8 - 30719999946x^4 + 17919999973) \partial^2 \\ & + 3x^2(102400000000x^{16} + 204800000000x^{12} + 289279999572x^8 - 350719999444x^4 + 30719999953) \partial \\ & - 3x(102400000000x^{16} + 204800000000x^{12} + 145919999796x^8 - 104959999828x^4 + 5119999993). \end{aligned}$$

This operator is an encoding of the algebraic function  $y = f(x)$  as a holonomic function.

In computer algebra, one represents a real algebraic number as a root of a polynomial with coefficients in  $\mathbb{Q}$ . However, this *minimal polynomial* does not specify the number uniquely. For that, one also needs an isolating interval or sign conditions on derivatives. The situation is analogous for encoding a holonomic function  $f$  in  $n$  variables. We specify  $f$  by a holonomic system of linear PDEs together with a list of initial conditions. The canonical holonomic representation is one possibility. Initial conditions such as (14.2) are designed to determine the function uniquely inside the linear space  $\text{Sol}(I)$ , where  $I \subseteq \text{Ann}_D(f)$ .

For instance, in Example 14.3, we would need three initial conditions to specify the function  $f(\mathbf{x})$  uniquely inside the 3-dimensional solution space to our operator  $P$ . We could fix the values at three distinct points, or we could fix the value and the first two derivatives at one special point.

To be more precise, we generalize the canonical representation (14.2) as follows. A *holonomic representation* of a function  $f$  is a holonomic  $D$ -ideal  $I \subseteq \text{ann}_D(f)$  together with a list of linear conditions that specify  $f$  uniquely inside the finite-dimensional solution space of holomorphic solutions. The existence of this representation makes  $f$  a *holonomic function*. The next example is meant to show the relevance of holonomic functions for metric algebraic geometry.

**Example 14.4 (The area of a TV screen)** Fix the quartic polynomial

$$q(x, y) = x^4 + y^4 + \frac{1}{100}xy - 1. \quad (14.3)$$

We are interested in the semi-algebraic set  $S = \{(x, y) \in \mathbb{R}^2 \mid q(x, y) \leq 0\}$ . This convex set is a slight modification of a set known in the optimization literature as “the TV screen”. Our aim is to compute the area of the semi-algebraic convex set  $S$  as accurately as is possible.

One can get a rough idea of the area of  $S$  by sampling. This is illustrated in Figure 14.2. From the polynomial  $q(x, y)$  we read off that  $S$  is contained in the square defined by  $-1.2 \leq x, y \leq 1.2$ . We sampled 10000 points uniformly from that square, and for each sample we checked the sign of  $q$ . Points inside  $S$  are drawn in blue and points outside  $S$  are drawn in pink. By multiplying the area  $(2.4)^2 = 5.76$  of the square with the fraction of the number of blue points among the samples, we learn that the area of the TV screen is approximately 3.7077.

We now compute the area more accurately using  $D$ -modules. Let  $\text{pr}: S \rightarrow \mathbb{R}$  be the projection on the  $x$ -coordinate, and write  $v(x) = \ell(\text{pr}^{-1}(x) \cap S)$  for the length of a fiber. This function is holonomic and it satisfies the third-order differential operator in Example 14.3.

The map  $\text{pr}$  has two branch points  $x_0 < x_1$ . They are the real roots of the resultant

$$\text{Res}_y(q, \partial q / \partial y) = 25600000000x^{12} - 76800000000x^8 + 76799999973x^4 - 25600000000. \quad (14.4)$$

These values can be written in radicals, but we take an accurate floating point representation:

$$x_1 = -x_0 = 1.000254465850258845478545766643566750080196276158976351763236\dots$$

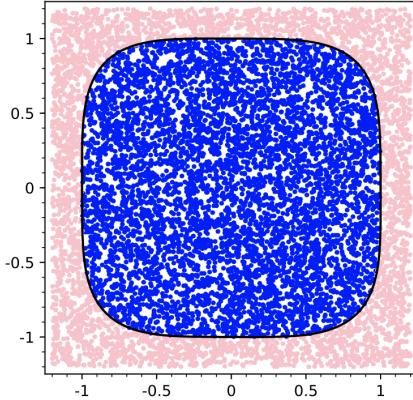


Fig. 14.2: The TV screen is the convex region consisting of the blue points.

The desired area equals  $\text{vol}(S) = w(x_1)$ , where  $w$  is the holonomic function

$$w(x) = \int_{x_0}^x v(t) dt.$$

One operator that annihilates  $w$  is  $P\partial$ , where  $P \in \text{ann}_D(v)$  is the third-order operator above. To get a holonomic representation of  $w$ , we also need some initial conditions. Clearly,  $w(x_0) = 0$ . Further initial conditions on  $w'$  are derived by evaluating  $v$  at other points. By plugging values for  $x$  into (14.3) and solving for  $y$ , we find  $w'(0) = 2$  and  $w'(\pm 1) = 1/\sqrt[3]{100}$ . Thus, we now have four linear constraints on our function  $w$ , albeit at different points.

Our goal is to determine a unique function  $w \in \text{Sol}(P\partial)$  by incorporating these four initial conditions, and then to evaluate  $w$  at  $x_1$ . To this end, we proceed as follows. Let  $x_{\text{ord}} \in \mathbb{R}$  be any point at which  $P\partial$  is not singular. Using the command `local_basis_expansion` that is built into the SAGE package `ore_algebra`, we compute a basis of local series solutions to  $P\partial$  at the point  $x_{\text{ord}}$ . Since that point is non-singular, that basis has the following form:

$$\begin{aligned} s_{x_{\text{ord}},0}(x) &= 1 + O((x - x_{\text{ord}})^4), \\ s_{x_{\text{ord}},1}(x) &= (x - x_{\text{ord}}) + O((x - x_{\text{ord}})^4), \\ s_{x_{\text{ord}},2}(x) &= (x - x_{\text{ord}})^2 + O((x - x_{\text{ord}})^4), \\ s_{x_{\text{ord}},3}(x) &= (x - x_{\text{ord}})^3 + O((x - x_{\text{ord}})^4). \end{aligned} \quad (14.5)$$

Locally at  $x_{\text{ord}}$ , our solution is given by a unique choice of four coefficients  $c_{x_{\text{ord}},i}$ , namely

$$w(x) = c_{x_{\text{ord}},0} \cdot s_{x_{\text{ord}},0}(x) + c_{x_{\text{ord}},1} \cdot s_{x_{\text{ord}},1}(x) + c_{x_{\text{ord}},2} \cdot s_{x_{\text{ord}},2}(x) + c_{x_{\text{ord}},3} \cdot s_{x_{\text{ord}},3}(x).$$

At a regular singular point  $x_{\text{rs}}$ , complex powers of  $x$  and  $\log(x)$  can appear in the local basis at  $x_{\text{rs}}$ . Any initial condition at that point determines a linear constraint on these coefficients. For instance,  $w'(0) = 2$  implies  $c_{0,1} = 2$ , and similarly for our initial conditions at  $-1, 1$  and  $x_0$ . One challenge is that the initial conditions pertain to different points. To address this, we calculate transition matrices that relate the basis (14.5) of series solutions at one point to the basis at another point. These are invertible  $4 \times 4$  matrices.

With the method described above, we find the basis of series solutions at  $x_1$ , along with a system of four linear constraints on the four coefficients  $c_{x_1,i}$ . These constraints are derived from the initial conditions at  $0, \pm 1$  and  $x_0$ , using the  $4 \times 4$  transition matrices. By solving these linear equations, we compute the

desired function value up to any desired precision:

$$w(x_1) = 3.708159944742162288348225561145865371243065819913934709438572\dots$$

In conclusion, this number is the area of the TV screen  $S$  defined by the polynomial  $q(x, y)$ .

Before computing an example in 3-space, let us first come back to properties of holonomic functions. Holonomic functions are very well-behaved with respect to many operations. They turn out to have remarkable closure properties. In the following, let  $f$  and  $g$  be functions in  $n$  variables  $\mathbf{x} = (x_1, \dots, x_n)$ .

**Proposition 14.5** *If  $f, g$  are holonomic functions, then both  $f + g$  and  $f \cdot g$  are holonomic.*

**Proof** For each index  $i \in \{1, 2, \dots, n\}$ , there exist non-zero operators  $P_i$  and  $Q_i$  in  $\mathbb{C}[\mathbf{x}]\langle\partial_i\rangle$  which satisfy  $P_i \bullet f = Q_i \bullet g = 0$ . Set  $n_i = \text{order}(P_i)$  and  $m_i = \text{order}(Q_i)$ . The  $\mathbb{C}(\mathbf{x})$ -linear span of the set  $\{\partial_i^k \bullet f\}_{k=0, \dots, n_i}$  has dimension  $\leq n_i$ . Similarly, the span of the set  $\{\partial_i^k \bullet g\}_{k=0, \dots, m_i}$  has dimension  $\leq m_i$ .

Now consider  $\partial_i^k \bullet (f + g) = \partial_i^k \bullet f + \partial_i^k \bullet g$ . The  $\mathbb{C}(\mathbf{x})$ -linear span of  $\{\partial_i^k \bullet (f + g)\}_{k=0, \dots, n_i+m_i}$  has dimension  $\leq n_i + m_i$ . Hence, there exists a non-zero operator  $S_i \in \mathbb{C}[\mathbf{x}]\langle\partial_i\rangle$ , such that  $S_i \bullet (f + g) = 0$ . Since this holds for all indices  $i$ , we conclude that the sum  $f + g$  is holonomic.

A similar proof works for the product  $f \cdot g$ . For each  $i \in \{1, 2, \dots, n\}$ , we now consider the set  $\{\partial_i^k \bullet (f \cdot g)\}_{k=0, 1, \dots, n_i m_i}$ . By applying Leibniz' rule for taking derivatives of a product, we find that the  $m_i n_i + 1$  generators are linear dependent over the rational function field  $\mathbb{C}(\mathbf{x})$ . Hence, there is a non-zero operator  $T_i \in \mathbb{C}[\mathbf{x}]\langle\partial_i\rangle$  such that  $T_i \bullet (f \cdot g) = 0$ . We conclude that the product  $f \cdot g$  is holonomic.  $\square$

The proof above gives a linear algebra method for computing an annihilating  $D$ -ideal  $I$  of finite holonomic rank for  $f + g$  (resp. of  $f \cdot g$ ), starting from such  $D$ -ideals for  $f$  and  $g$ . The following example, similar to one in [184, Section 4.1], illustrates Proposition 14.5.

**Example 14.6 ( $n = 1$ )** Consider the functions  $f(x) = \exp(x)$  and  $g(x) = \exp(-x^2)$ . Their canonical holonomic representations are  $I_f = \langle\partial - 1\rangle$  with  $f(0) = 1$  and  $I_g = \langle\partial + 2x\rangle$  with  $g(0) = 1$ . We are interested in the function  $h = f + g$ . Its first partial derivatives are

$$\begin{pmatrix} h \\ \partial \bullet h \\ \partial^2 \bullet h \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -2x \\ 1 & 4x^2 - 2 \end{pmatrix} \cdot \begin{pmatrix} f \\ g \end{pmatrix}.$$

By computing the left kernel of this  $3 \times 2$ -matrix, we find that  $h = f + g$  is annihilated by

$$I_h = \langle(2x + 1)\partial^2 + (4x^2 - 3)\partial - 4x^2 - 2x + 2\rangle, \quad \text{with } h(0) = 2, h'(0) = 1.$$

For the product  $j = f \cdot g$  we have  $j' = f'g + fg' = f \cdot g + f \cdot (-2xg) = (1 - 2x)j$ , so the canonical holonomic representation of  $j$  is the  $D$ -ideal  $I_j = \langle\partial + 2x - 1\rangle$  with  $j(0) = 1$ .

**Proposition 14.7** *Let  $f$  be holonomic in  $n$  variables and  $m < n$ . Then the restriction of  $f$  to the coordinate subspace  $\{x_{m+1} = \dots = x_n = 0\}$  is a holonomic function in the first  $m$  variables  $x_1, \dots, x_m$ .*

**Proof** For  $i \in \{m + 1, \dots, n\}$ , we consider the right ideal  $x_i D$  in the Weyl algebra  $D$ . This ideal is a left module over  $D_m = \mathbb{C}\langle x_1, \dots, x_m, \partial_1, \dots, \partial_m \rangle$ . The sum of these ideals with  $\text{Ann}_D(f)$  is hence a left  $D_m$ -module. Its intersection with  $D_m$  is called the *restriction ideal*:

$$(\text{Ann}_D(f) + x_{m+1}D + \dots + x_nD) \cap D_m. \tag{14.6}$$

By [160, Prop. 5.2.4], this  $D_m$ -ideal is holonomic and it annihilates  $f(x_1, \dots, x_m, 0, \dots, 0)$ .  $\square$

**Proposition 14.8** *The partial derivatives of a holonomic function are holonomic functions.*

**Proof** Let  $f$  be holonomic and  $P_i \in \mathbb{C}[x]\langle\partial_i\rangle \setminus \{0\}$  with  $P_i \bullet f = 0$  for all  $i$ . We can write  $P_i$  as  $P_i = \tilde{P}_i \partial_i + a_i(x)$ , where  $a_i \in \mathbb{C}[x]$ . If  $a_i = 0$ , then  $\tilde{P}_i \bullet \frac{\partial f}{\partial x_i} = 0$  and we are done. Assume  $a_i \neq 0$ . Since both  $a_i$  and  $f$  are holonomic, by Proposition 14.5, there is a non-zero linear operator  $Q_i \in \mathbb{C}[x]\langle\partial_i\rangle$  such that  $Q_i \bullet (a_i \cdot f) = 0$ . Then  $Q_i \tilde{P}_i$  annihilates  $\partial f / \partial x_i$ .  $\square$

A key insight from the theory of  $D$ -modules (see [160, Section 5.5]) is that integration is dual, in the sense of the Fourier transform, to restriction. Here is the dual to Proposition 14.7.

**Proposition 14.9** *Let  $f: \mathbb{R}^n \rightarrow \mathbb{C}$  be a holonomic function. Then the definite integral*

$$F(x_1, \dots, x_{n-1}) = \int_a^b f(x_1, \dots, x_{n-1}, x_n) dx_n$$

*is a holonomic function in  $n - 1$  variables, assuming the integral converges.*

By dualizing (14.6), we obtain the following  $D_m$ -ideal, known as the *integration ideal*:

$$(\text{Ann}_D(f) + \partial_{m+1} D + \dots + \partial_n D) \cap D_m \quad \text{for } m < n.$$

The expression is dual to the restriction ideal (14.6) under the Fourier transform. This exchanges  $x_i$  and  $\partial_i$ . If  $m = n - 1$  then the integration ideal annihilates the holonomic function  $F$  above.

Equipped with our tools for holonomic functions, we now return to computing volumes of compact semi-algebraic sets. We follow the work of Lairez, Mezzarobba and Safey El Din in [119]. They compute this volume by deriving a differential operator that encodes the period of a certain integral [118]. Here is the definition. Let  $R(t, x_1, \dots, x_n)$  be a rational function and consider the formal period integral

$$\oint R(t, x_1, \dots, x_n) dx_1 \cdots dx_n. \quad (14.7)$$

Fix an open subset  $\Omega$  of either  $\mathbb{R}$  or  $\mathbb{C}$ . An analytic function  $\phi: \Omega \rightarrow \mathbb{C}$  is a *period* of the integral (14.7) if, for any  $s \in \Omega$ , there exists a neighborhood  $\Omega' \subseteq \Omega$  of  $s$  and an  $n$ -cycle  $\gamma \subset \mathbb{C}^n$  with the following property. For all  $t \in \Omega'$ ,  $\gamma$  is disjoint from the poles of  $R_t := R(t, \bullet)$  and

$$\phi(t) = \int_{\gamma} R(t, x_1, \dots, x_n) dx_1 \cdots dx_n. \quad (14.8)$$

If this holds, then there exists an operator  $P \in D \setminus \{0\}$  of the Fuchsian class annihilating  $\phi(t)$ .

Let  $S = \{f \leq 0\} \subset \mathbb{R}^n$  be a compact basic semi-algebraic set, defined by a polynomial  $f \in \mathbb{Q}[x_1, \dots, x_n]$ . Let  $\text{pr}: \mathbb{R}^n \rightarrow \mathbb{R}$  denote the projection on the first coordinate. The set of *branch points* of  $\text{pr}$  is the following subset of the real line, which is assumed to be finite:

$$\Sigma_f = \{p \in \mathbb{R} \mid \exists x = (x_2, \dots, x_n) \in \mathbb{R}^{n-1} : f(p, x) = 0 \text{ and } \frac{\partial f}{\partial x_i}(p, x) = 0 \text{ for } i = 2, \dots, n\}.$$

The polynomial in the unknown  $p$  that defines  $\Sigma_f$  is obtained by eliminating  $x_2, \dots, x_n$ . It can be represented as a multivariate resultant, generalizing the Sylvester resultant in (14.4).

Fix an open interval  $I$  in  $\mathbb{R}$  with  $I \cap \Sigma_f = \emptyset$ . For any  $x_1 \in I$ , the set  $S_{x_1} := \text{pr}^{-1}(x_1) \cap S$  is compact and semi-algebraic in  $(n - 1)$ -space. We are interested in its volume. By [119, Theorem 9], the function  $v: I \rightarrow \mathbb{R}$ ,  $x_1 \mapsto \text{vol}_{n-1}(S_{x_1})$  is a period of the rational integral

$$\frac{1}{2\pi i} \oint \frac{x_2}{f(x_1, x_2, \dots, x_n)} \frac{\partial f(x_1, x_2, \dots, x_n)}{\partial x_2} dx_2 \cdots dx_n. \quad (14.9)$$

Let  $e_1 < e_2 < \dots < e_K$  be the branch points in  $\Sigma_f$  and set  $e_0 = -\infty$  and  $e_{K+1} = \infty$ . This specifies the pairwise disjoint open intervals  $I_k = (e_k, e_{k+1})$ . They satisfy  $\mathbb{R} \setminus \Sigma_f = \bigcup_{k=0}^K I_k$ . Fix the holonomic functions  $w_k(t) = \int_{e_k}^t v(x_1) dx_1$ . The volume of  $S$  then is obtained as

$$\text{vol}_n(S) = \int_{e_1}^{e_K} v(x_1) dx_1 = \sum_{k=1}^{K-1} w_k(e_{k+1}).$$

How does one compute such an expression? As a period of the rational integral (14.9), the volume  $v$  is a holonomic function on each interval  $I_k$ . A key step is to find an operator  $P \in D_1$  that annihilates  $v|_{I_k}$  for all  $k$ . Then the product  $P\partial$  annihilates the functions  $w_k(x_1)$  for all  $k$ . By imposing sufficiently many initial conditions, we can reconstruct the functions  $w_k$  from the operator  $P\partial$ . One initial condition that comes for free for each  $k$  is  $w_k(e_k) = 0$ .

The differential operator  $P$  is known as the *Picard–Fuchs equation* of the period in question. The following software packages can be used to compute such Picard–Fuchs equations:

- `HolonomicFunctions` by C. Koutschan in **Mathematica**,
- `ore_algebra` by M. Kauers in **SAGE**,
- `periods` by P. Lairez in **MAGMA**,
- `Ore_Algebra` by F. Chyzak in **Maple**.

We now apply this to compute volumes. Starting from the polynomial  $f$ , we compute the Picard–Fuchs operator  $P \in D_1$  along with sufficiently many compatible initial conditions. For each interval  $I_k$ , where  $k = 1, \dots, K-1$ , we perform the following steps, here described for the `ore_algebra` package in **SAGE**:

1. Using the command `local_basis_expansion`, compute a local basis of series solutions for the linear differential operator  $P\partial$  at various points in  $[e_k, e_{k+1}]$ .
2. Using the command `op.numerical_transition_matrix`, numerically compute a transition matrix for the series solution basis from one point to another one.
3. From the initial conditions construct linear relations between the coefficients in the local basis extensions. Using step (ii), transfer them to the branch point  $e_{k+1}$ .
4. Plug in to the local basis extension at  $e_{k+1}$  and thus evaluate the volume of  $S \cap \text{pr}^{-1}(I_k)$ .

We illustrate this recipe by computing the volume of a convex body in 3-space, shown in Figure 14.3.

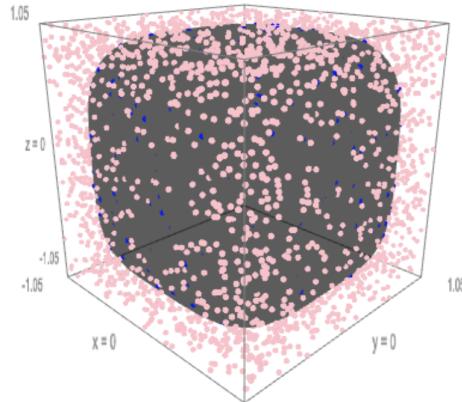


Fig. 14.3: The quartic bounds the convex region consisting of the gray points.

**Example 14.10 (Quartic surface)** Fix the quartic polynomial

$$f(x, y, z) = x^4 + y^4 + z^4 + \frac{x^3y}{20} - \frac{xyz}{20} - \frac{yz}{100} + \frac{z^2}{50} - 1, \quad (14.10)$$

and consider the set  $S = \{(x, y, z) \in \mathbb{R}^3 \mid f(x, y, z) \leq 0\}$ . Our aim is to compute  $\text{vol}_3(S)$ .

As in Example 14.4 with the TV screen, we can get a rough idea of the volume of  $S$  by sampling. This is illustrated in Figure 14.3. Our set  $S$  is compact, convex, and contained in the cube defined by  $-1.05 \leq x, y, z \leq 1.05$ . We sampled 10000 points uniformly from that cube. For each sample we checked the sign of  $f(x, y, z)$ . By multiplying the volume  $(2.1)^3 = 9.261$  of the cube by the fraction of the number of gray points and the number of sampled points, SAGE found within few seconds that  $\text{vol}(S) \approx 6.4771$ . In order to obtain a higher precision, we now compute the volume of our set  $S$  with help of  $D$ -modules.

Let  $\text{pr}: \mathbb{R}^3 \rightarrow \mathbb{R}$  be the projection onto the  $x$ -coordinate. Let  $v(x) = \text{vol}_2(\text{pr}^{-1}(x) \cap S)$  denote the area of the fiber over any point  $x$  in  $\mathbb{R}$ . We write  $e_1 < e_2$  for the two branch points of the map  $\text{pr}$  restricted to the quartic surface  $\{f = 0\}$ . They can be computed with resultants. The projection has 36 complex branch points. The first two of them are real and therefore are the branch points of  $\text{pr}$ . We obtain  $e_1 \approx -1.0023512$  and  $e_2 \approx 1.0024985$ . By [119, Theorem 9], the area function  $v(x)$  is a period of the rational integral

$$\frac{1}{2\pi i} \oint \frac{y}{f(x, y, z)} \frac{\partial f(x, y, z)}{\partial y} dy dz.$$

We set  $w(t) = \int_{e_1}^t v(x) dx$ . The desired 3-dimensional volume equals  $\text{vol}_3(S) = w(e_2)$ .

Using Lairez' implementation periods in MAGMA, we compute a differential operator  $P$  of order eight that annihilates  $v(x)$ . Again,  $P\partial$  then annihilates  $w(x)$ . One initial condition is  $w(e_1) = 0$ . We obtain eight further initial conditions  $w'(x) = \text{vol}_2(S_x)$  for points  $x \in (e_1, e_2)$  by running the same algorithm for the 2-dimensional semi-algebraic slices  $S_x = \text{pr}^{-1}(x) \cap S$ . In other words, we make eight subroutine calls to an area measurement as in Example 14.4. From these nine initial conditions we derive linear relations of the coefficients in the local basis expansion at  $e_2$ . These computations are run in SAGE as described in steps (i), (ii), (iii) and (iv) above. We find the approximate volume of our convex body  $S$  to be

$$\begin{aligned} &\approx 6.438832480572893544740733895969956188958420889235116976266328923128826 \\ &9155273887642162091495583989038294311376088934526903525560097601024171 \\ &190804769405534826558114212766135380613959757935305271022089419155701 \\ &52158647017087400219438452914068685622775954171509711339913473405961 \\ &7632892206072085516332397969163383760070738760107318247752061504714 \\ &367250460900923409066377732273390396822296235214963623286613117557 \\ &930687544148360721225681053481178760058264738867105810326818911 \\ &578448323758536767168707442532146029753762594261578920477859. \end{aligned}$$

This numerical value is guaranteed to be accurate up to 550 digits.

### 14.3 Lasserre's Method

In this section we present the second method for computing volumes, based on semidefinite programming. This was developed by Lasserre and his collaborators. See [90, 175, 176] and references therein. We consider an inclusion of semialgebraic sets  $K \subset B \subset \mathbb{R}^n$ , where  $K$  and  $B$  are compact. Here  $B$  is a set that serves as a bounding box, like  $B = [-1, 1]^n$ . We assume that the moments of Lebesgue measure on  $B$  are known or easy-to-compute. In other words, we assume that we have access to the values of the integrals

$$\beta_{\mathbf{u}} = \int_B \mathbf{x}^{\mathbf{u}} d\mathbf{x} = \int_B x_1^{u_1} x_2^{u_2} \cdots x_n^{u_n} dx_1 dx_2 \cdots dx_n \quad \text{for } \mathbf{u} \in \mathbb{N}^n.$$

The moments  $m_{\mathbf{u}}$  of Lebesgue measure on  $K$  are unknown. These will be our decision variables:

$$m_{\mathbf{u}} = \int_K \mathbf{x}^{\mathbf{u}} d\mathbf{x} = \int_K x_1^{u_1} x_2^{u_2} \cdots x_n^{u_n} dx_1 dx_2 \cdots dx_n \quad \text{for } \mathbf{u} \in \mathbb{N}^n. \quad (14.11)$$

Our aim is to compute  $m_0 = \text{vol}(K)$ . The idea is to use the following infinite-dimensional linear program: *Maximize the integral  $\int d\mu$ , where  $\mu$  and  $\hat{\mu}$  range over measures on  $\mathbb{R}^n$ , where  $\mu$  is supported on  $K$ ,  $\hat{\mu}$  is supported on  $B$ , and the sum  $\mu + \hat{\mu}$  is Lebesgue measure on  $B$ .*

The unique optimal solution  $(\mu, \hat{\mu})$  to this linear program can be characterized as follows:  $\mu^*$  is Lebesgue measure on  $K$ ,  $\hat{\mu}^*$  is Lebesgue measure on  $B \setminus K$ , and the optimal value is  $\text{vol}(K) = \int d\mu^*$ . This is described in [176, equation (1)]. The linear programming (LP) dual is given in [176, equation (2)].

We can express our linear program in terms of the moment sequences  $\mathbf{m} = (m_{\mathbf{u}})$  and  $\hat{\mathbf{m}} = (\hat{m}_{\mathbf{u}})$  of the two unknown measures  $\mu$  and  $\hat{\mu}$ . Namely, we paraphrase: *Maximize  $m_0$  subject to  $m_{\mathbf{u}} + \hat{m}_{\mathbf{u}} = \beta_{\mathbf{u}}$  for all  $\mathbf{u} \in \mathbb{N}^d$ , where  $\mathbf{m}$  and  $\hat{\mathbf{m}}$  are valid moment sequences of measures on  $\mathbb{R}^n$ , with  $m$  supported on  $K$ .* This brings us to the moment problem, which is the question how to characterize valid moment sequences. This is a problem with a long history in mathematics, and an exact characterization is very difficult. However, in recent years, it has been realized that there are effective necessary conditions. These involve semidefinite programming formulations in finite dimensions, which are built via the *localizing matrices* we now define.

Let  $K = \{\mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) \geq 0\}$  be defined by a single polynomial  $f = \sum_{\mathbf{w}} c_{\mathbf{w}} \mathbf{x}^{\mathbf{w}}$  in  $n$  variables. Fix an integer  $d$  that exceeds the degree of  $f$ . We shall construct three symmetric matrices of format  $\binom{n+d}{d} \times \binom{n+d}{d}$  whose entries are linear in the decision variables. The rows and columns of our matrices are indexed by elements  $\mathbf{u} \in \mathbb{N}^n$  with  $|\mathbf{u}| = u_1 + \cdots + u_n$  at most  $d$ . These correspond to monomials  $\mathbf{x}^{\mathbf{u}}$  of degree  $\leq d$ .

Our first matrix  $M_d(\mathbf{m})$  has the entry  $m_{\mathbf{u}+\mathbf{v}}$  in row  $\mathbf{u}$  and column  $\mathbf{v}$ . Our second matrix  $M_d(\hat{\mathbf{m}})$  has the entry  $\hat{m}_{\mathbf{u}+\mathbf{v}}$  in row  $\mathbf{u}$  and column  $\mathbf{v}$ . And, finally, our third matrix  $M_d(f\mathbf{m})$  has the entry  $\sum_{\mathbf{w}} c_{\mathbf{w}} m_{\mathbf{u}+\mathbf{v}+\mathbf{w}}$  in row  $\mathbf{u}$  and column  $\mathbf{v}$ . We consider the following semidefinite program:

$$\begin{aligned} & \text{Maximize } m_0 \text{ subject to } m_{\mathbf{u}} + \hat{m}_{\mathbf{u}} = \beta_{\mathbf{u}} \text{ for all } \mathbf{u} \in \mathbb{N}^d \text{ with } |\mathbf{u}| \leq d, \text{ where} \\ & \text{the symmetric matrices } M_d(m), M_d(\hat{m}) \text{ and } M_d(f\mathbf{m}) \text{ are positive semidefinite.} \end{aligned} \quad (14.12)$$

Here, the third matrix is usually replaced by  $M_{d'}(f\mathbf{m})$  where  $d' = d - \lceil \deg(f)/2 \rceil$ . The objective function value depends on  $d$ , and it decreases as  $d$  increases. The limit for  $d \rightarrow \infty$  is equal to the volume of  $X$ . Indeed, this sequence of SDP problems is an approximation to the infinite-dimensional linear programming problem above. The convergence property was proved in [90].

The remainder of this section will demonstrate how one can solve (14.12) in practice. It is based on [90, 175, 176], and we discuss an implementation in **Mathematica**. This material was developed by Chiara Meroni, and we are very grateful to her for allowing us to include it in these lecture notes.

Our point of departure is the following question: given a sequence of real numbers  $\mathbf{m} = (m_{\alpha})_{\alpha}$ , does there exist a set  $S$  and a measure  $\mu_S$  supported on  $S$  such that (14.11) holds? Given  $d \in \mathbb{N}$ , denote by  $\mathbb{N}_d^n$  the set of multiindices  $\alpha \in \mathbb{N}^n$  such that  $|\alpha| = \alpha_1 + \dots + \alpha_n \leq d$ . Fix a semialgebraic set  $K$  as above, let  $r = \lceil \frac{\deg f}{2} \rceil$ , and consider an arbitrary sequence of real numbers  $\mathbf{m} = (m_{\alpha})_{\alpha}$  that is indexed by  $\mathbb{N}_d^n$ .

The *moment matrix* and the *localizing matrix* associated with this sequence are respectively

$$M_d(\mathbf{m}) = \left( m_{\alpha+\beta} \right)_{\alpha, \beta \in \mathbb{N}_d^n}, \quad M_{d-r}(f\mathbf{m}) = \left( \sum_{\mathbf{w} \in W} c_{\mathbf{w}} m_{\mathbf{w}+\alpha+\beta} \right)_{\alpha, \beta \in \mathbb{N}_{d-r}^n}. \quad (14.13)$$

The moment matrix has size  $\binom{n+d}{d} \times \binom{n+d}{d}$  whereas the localizing matrix has size  $\binom{n+d-r}{d-r} \times \binom{n+d-r}{d-r}$ . A necessary condition for a sequence  $\mathbf{m} = (m_{\alpha})_{\alpha}$  to have a representing measure supported on  $K$  is that for

every  $d \in \mathbb{N}$  the matrix inequalities  $M_d(\mathbf{m}) \succeq 0$  and  $M_{d-r}(f\mathbf{m}) \succeq 0$  hold. This result is a formulation of Putinar's Positivstellensatz [90, Theorem 2.2]. In particular, the positive definiteness of the moment matrix is a necessary condition for  $\mathbf{m}$  to have a representing measure; the inequality with the localizing matrix forces the support of the representing measure to be contained in the superlevel set  $\{f(\mathbf{x}) \geq 0\}$ , namely  $K$ .

**Example 14.11** As a sanity check, consider the disc  $K = \{(x, y) \in \mathbb{R}^2 \mid f = 1 - x^2 - y^2 \geq 0\}$ . Its moments are

$$m_{(\alpha_1, \alpha_2)} = ((-1)^{\alpha_1} + 1)((-1)^{\alpha_2} + 1) \frac{\Gamma\left(\frac{\alpha_1+1}{2}\right)\Gamma\left(\frac{\alpha_2+1}{2}\right)}{4\Gamma\left(\frac{1}{2}(\alpha_1 + \alpha_2 + 4)\right)},$$

where  $\Gamma$  denotes the usual Gamma function. For  $d = 3$ , the moment and localizing matrices in (14.13) are

$$M_3(\mathbf{m}) = \begin{pmatrix} \pi & 0 & \frac{\pi}{4} & 0 & 0 & 0 & 0 & \frac{\pi}{4} & 0 & 0 \\ 0 & \frac{\pi}{4} & 0 & \frac{\pi}{8} & 0 & 0 & 0 & 0 & \frac{\pi}{24} & 0 \\ \frac{\pi}{4} & 0 & \frac{\pi}{8} & 0 & 0 & 0 & 0 & \frac{\pi}{24} & 0 & 0 \\ 0 & \frac{\pi}{8} & 0 & \frac{5\pi}{64} & 0 & 0 & 0 & 0 & \frac{\pi}{64} & 0 \\ 0 & 0 & 0 & \frac{\pi}{24} & 0 & \frac{\pi}{24} & 0 & 0 & 0 & \frac{\pi}{8} \\ 0 & 0 & 0 & 0 & \frac{\pi}{24} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{\pi}{64} & 0 & 0 & 0 & \frac{\pi}{64} \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{\pi}{64} & 0 & 0 & 0 \\ \frac{\pi}{4} & 0 & \frac{\pi}{24} & 0 & 0 & 0 & \frac{\pi}{8} & 0 & 0 & 0 \\ 0 & \frac{\pi}{24} & 0 & \frac{\pi}{64} & 0 & 0 & 0 & \frac{\pi}{64} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{\pi}{8} & 0 & 0 & \frac{\pi}{64} & 0 & 0 \end{pmatrix}, \quad M_2(f\mathbf{m}) = \begin{pmatrix} \frac{\pi}{2} & 0 & \frac{\pi}{12} & 0 & 0 & \frac{\pi}{12} \\ 0 & \frac{\pi}{12} & 0 & 0 & 0 & 0 \\ \frac{\pi}{12} & 0 & \frac{\pi}{32} & 0 & 0 & \frac{\pi}{96} \\ 0 & 0 & 0 & \frac{\pi}{12} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{\pi}{96} & 0 \\ \frac{\pi}{12} & 0 & \frac{\pi}{96} & 0 & 0 & \frac{\pi}{32} \end{pmatrix},$$

These two matrices are symmetric and positive definite.

We consider the infinite-dimensional *linear program* on measures whose optimal value is the volume of  $K \subset B$ . The program was stated above. We use the formulation in [90, Equation 3.1] and [176, Equation 1]:

$$\begin{aligned} P : \quad & \max_{\mu_K, \mu_{B \setminus K}} \int d\mu_K \\ & \text{s.t. } \mu_K + \mu_{B \setminus K} = \mu_B^*. \end{aligned} \tag{14.14}$$

Here  $\mu_S$  is a positive finite Borel measure supported on  $S$ , and  $\mu_B^*$  is the Lebesgue measure on  $B$ . The adjective “infinite-dimensional” refers to the fact that we are optimizing over a set of measures, which is uncountable. Based on the theory of dual Banach spaces, one can talk about dual convex bodies, and construct a duality theory for LP. In our case, the dual to the space of positive finite Borel measures is the set of positive continuous functions. This observation leads to the definition of an LP dual to  $P$ :

$$\begin{aligned} P^* : \quad & \inf_{\gamma} \int \gamma d\mu_B^* \\ & \text{s.t. } \gamma \geq \mathbf{1}_K, \end{aligned} \tag{14.15}$$

where  $\gamma$  is a positive continuous function on  $B$  and  $\mathbf{1}_K$  is the indicator function of  $K$ . There is no duality gap between  $P$  and  $P^*$ , i.e. the optimal values of (14.14) and (14.15) coincide. Note that the optimal value of  $P^*$  is an infimum and not a minimum, since we are approximating the *discontinuous* indicator function  $\mathbf{1}_K$  using continuous functions. This detail explains the slow rate of approximation of the basic method.

The infinite-dimensional LP can be approximated by a hierarchy of finite-dimensional *semidefinite programs* [120]. The optimal values of the hierarchy converge monotonically to the optimal value of the LP [90, Theorem 3.2]. There is again a primal and dual version of the SDP. In our setting, the primal is

$$\begin{aligned} P_d : \quad & \max_{\mathbf{m}, \widehat{\mathbf{m}}} m_0 \\ \text{s.t. } & \mathbf{m} + \widehat{\mathbf{m}} = \mathbf{b}, \quad M_d(\mathbf{m}) \geq 0, \quad M_d(\widehat{\mathbf{m}}) \geq 0, \quad M_{d-r}(f\mathbf{m}) \geq 0. \end{aligned} \quad (14.16)$$

Here  $\mathbf{m} = (m_\alpha)_{\alpha \in \mathbb{N}_{2d}^n}$ ,  $\widehat{\mathbf{m}} = (\widehat{m}_\alpha)_{\alpha \in \mathbb{N}_{2d}^n}$ , and  $\mathbf{b}$  contains the moments of  $B$  indexed by  $\mathbb{N}_{2d}^n$ . This formulation is [176, Equation 3]. The optimal value of  $P_d$  is an upper bound for  $\text{vol}(K)$ , since we are optimizing over a larger set. The dual SDP is [90, Equation 3.6], which is formulated using sums of squares of polynomials. The authors of [90, 175, 176] implemented the SDPs using **GloptiPoly** MATLAB. Our computations in the next examples are performed in **Mathematica**. We are going to include the linear condition  $\mathbf{m} + \widehat{\mathbf{m}} = \mathbf{b}$  inside the condition on the moment matrix of  $\widehat{\mathbf{m}}$ , by imposing directly that  $M_d(\mathbf{b} - \mathbf{m}) \geq 0$ .

**Example 14.12 (TV screen)** Fix  $K_1 = \{(x, y) \in [-1.2, 1.2]^2 \mid f_1(x, y) \geq 0\} \subset \mathbb{R}^2$  where  $f_1 = -q$  is the quartic in (14.3). This convex set shown in Figures 14.2 and 14.4. Recall that  $\text{vol}(K_1) = 3.7081599447\dots$

Let us now try the SDP formulation above, with  $d = 10$ . The moment matrices  $M_{10}(\mathbf{m})$  and  $M_{10}(\mathbf{b} - \mathbf{m})$  have format  $66 \times 66$ . For instance, the second matrix looks like

$$M_{10}(\mathbf{b} - \mathbf{m}) = \begin{pmatrix} 4-m_{(0,0)} & -m_{(0,1)} & \frac{4}{3}-m_{(0,2)} & -m_{(0,3)} & \cdots \\ -m_{(0,1)} & \frac{4}{3}-m_{(0,2)} & -m_{(0,3)} & \frac{4}{5}-m_{(0,4)} & \cdots \\ \frac{4}{3}-m_{(0,2)} & -m_{(0,3)} & \frac{4}{5}-m_{(0,4)} & -m_{(0,5)} & \cdots \\ -m_{(0,3)} & \frac{4}{5}-m_{(0,4)} & -m_{(0,5)} & \frac{4}{7}-m_{(0,6)} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

The localizing matrix  $M_8(f_1\mathbf{m})$  has format  $45 \times 45$ . Its  $(\alpha, \beta)$  entry equals

$$m_{\alpha+\beta} - m_{(4,0)+\alpha+\beta} - m_{(0,4)+\alpha+\beta} - \frac{1}{100}m_{(1,1)+\alpha+\beta}.$$

The optimal value of the semidefinite program  $P_{10}$  is  $4.4644647361\dots$ , the optimal value of  $P_{14}$  is  $4.3679560947\dots$ , and for  $P_{18}$  we get  $4.3241824171\dots$ . These numbers are upper bounds for the actual volume, as predicted. However, these bounds are still far from the truth.

**Example 14.13 (Elliptope)** Set  $f_2(x, y) = 1 - x^2 - y^2 - z^2 + 2xyz$ . This defines the elliptope  $K_2 = \{x, y \in [-1, 1]^3 \mid f_2(x, y) \geq 0\} \subset \mathbb{R}^3$ , which is shown in Figures 14.1 and 14.4. We already know  $\text{vol } K_2 = \frac{\pi^2}{2} = 4.934802202\dots$ . The upper bounds computing from the semidefinite program for  $d = 4, 8, 12$  are respectively  $7.3254012963\dots$ ,  $6.6182632506\dots$ , and  $6.303035372\dots$ . This is still pretty bad.

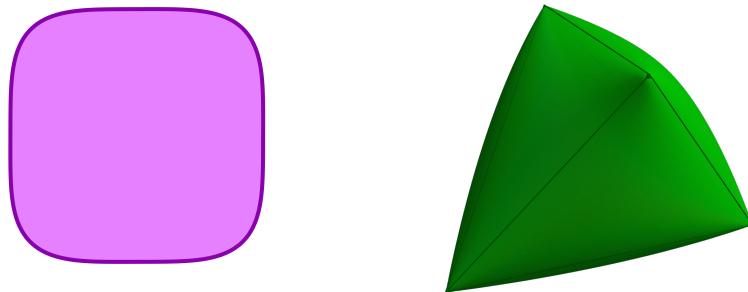


Fig. 14.4: Left: the TV screen from Example 14.12. Right: the elliptope from Example 14.13.

Examples 14.12 and 14.13 suggest that the convergence of the SDP approximation is quite slow. To improve the convergence, one uses the method of *Stokes constraints*. This was introduced and analyzed

in [175, 176] and we shall now explain it. In the infinite-dimensional linear program  $P^*$  (and in the SDP hierarchy) we aim to approximate a piecewise-differentiable function,  $\mathbf{1}_K$  with continuous functions (respectively, polynomials). This produces the well-known *Gibbs effect*, creating many oscillations near the boundary of  $K$  in the polynomial solutions of the SDP. To remedy this, we add linear constraint that do not modify the infinite-dimensional LP problem but add more information to the finite-dimensional SDP. One concrete way to do this uses Stokes' theorem and the fact that  $f$  vanishes on the boundary  $\partial K$  of  $K$ .

Let  $U$  be an open region in  $\mathbb{R}^n$  such that the Euclidean closure of  $U$  is our semialgebraic set  $K$ . Since  $\partial K$  is smooth almost everywhere, the classical Stokes Theorem applies. This theorem states that

$$\int_{\partial K} \omega = \int_K d\omega$$

for any  $(n - 1)$ -differential form  $\omega$  on  $\mathbb{R}^n$ . One consequence of Stokes' Theorem is *Gauss' formula*

$$\int_{\partial K} V(\mathbf{x}) \cdot \hat{n}(\mathbf{x}) d\mathcal{H}^{n-1}(\mathbf{x}) = \int_K \operatorname{div} V(\mathbf{x}) d\mathbf{x}.$$

Here  $V(\mathbf{x})$  is a vector field,  $\operatorname{div}$  denotes divergence,  $\hat{n}(\mathbf{x})$  is the exterior normal vector at  $\mathbf{x} \in \partial K$ , and  $\mathcal{H}^{n-1}$  is  $(n - 1)$ -dimensional Hausdorff measure. If the vector field is a scalar field times a constant vector  $\mathbf{c} \in \mathbb{R}^n$ , say  $V(\mathbf{x}) = v(\mathbf{x})\mathbf{c}$ , then we obtain the following equations:

$$\mathbf{c} \cdot \left( \int_{\partial K} v(\mathbf{x}) \hat{n}(\mathbf{x}) d\mathcal{H}^{n-1}(\mathbf{x}) \right) = \int_K \operatorname{div}(v(\mathbf{x})\mathbf{c}) d\mathbf{x} = \mathbf{c} \cdot \left( \int_K \nabla v(\mathbf{x}) d\mathbf{x} \right).$$

This holds because  $\operatorname{div}(v(\mathbf{x})\mathbf{c}) = \nabla v(\mathbf{x}) \cdot \mathbf{c} + v(\mathbf{x}) \operatorname{div} \mathbf{c}$  and the divergence of a constant vector is zero. Since this identity holds for every  $\mathbf{c} \in \mathbb{R}^n$ , we have

$$\int_{\partial K} v(\mathbf{x}) \hat{n}(\mathbf{x}) d\mathcal{H}^{n-1}(\mathbf{x}) = \int_K \nabla v(\mathbf{x}) d\mathbf{x}. \quad (14.17)$$

If  $v = 0$  on  $\partial K$ , then the left hand side of (14.17) is zero. This condition can be expressed in terms of measures and distributions, and added to (14.14) and (14.15) as in [176, Equation 17 and Remark 3].

In the setting of our SDP hierarchy, the Stokes constraints are written as follows. Let  $v(\mathbf{x}) = f(\mathbf{x})\mathbf{x}^\alpha$  for any multiindex  $\alpha \in \mathbb{N}^n$  with  $|\alpha| \leq d + 1 - \deg f$ . Then we require

$$\nabla(f(\mathbf{x})\mathbf{x}^\alpha)|_{\mathbf{x}^\beta \rightarrow m_\beta} = 0.$$

We now replace each monomial by its moment. This yields  $n$  new linear conditions for each  $\alpha$  as above.

**Example 14.14** For the SDP in Examples 14.12 and 14.13, the Stokes constraints for a given  $\alpha$  are:

$K_1$  :

$$\begin{aligned} \alpha_1 m_{\alpha+(-1,0)} - (\alpha_1 + 4)m_{\alpha+(3,0)} - \alpha_1 m_{\alpha+(-1,4)} - \frac{\alpha_1 + 1}{100}m_{\alpha+(0,1)} &= 0, \\ \alpha_2 m_{\alpha+(0,-1)} - \alpha_2 m_{\alpha+(4,-1)} - (\alpha_2 + 4)m_{\alpha+(0,3)} - \frac{\alpha_2 + 1}{100}m_{\alpha+(1,0)} &= 0, \end{aligned}$$

$K_2$  :

$$\begin{aligned} \alpha_1 m_{\alpha+(-1,0,0)} - (\alpha_1 + 2)m_{\alpha+(1,0,0)} - \alpha_1 m_{\alpha+(-1,2,0)} - \alpha_1 m_{\alpha+(-1,0,2)} + 2(\alpha_1 + 1)m_{\alpha+(0,1,1)} &= 0, \\ \alpha_2 m_{\alpha+(0,-1,0)} - \alpha_2 m_{\alpha+(2,-1,0)} - (\alpha_2 + 2)m_{\alpha+(0,1,0)} - \alpha_2 m_{\alpha+(0,-1,2)} + 2(\alpha_2 + 1)m_{\alpha+(1,0,1)} &= 0, \\ \alpha_3 m_{\alpha+(0,0,-1)} - \alpha_3 m_{\alpha+(2,0,-1)} - \alpha_3 m_{\alpha+(0,2,-1)} - (\alpha_3 + 2)m_{\alpha+(0,0,1)} + 2(\alpha_3 + 1)m_{\alpha+(1,1,0)} &= 0. \end{aligned}$$

Table 14.1 compares the optimal values of the SDP (14.13) with and without Stokes constraints.

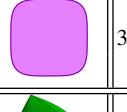
| $K$   | Volume      | $d$ | without Stokes |            | with Stokes |             |
|---|-------------|-----|----------------|------------|-------------|-------------|
|   |             |     | $\max P_d$     | time       | $\max P_d$  | time        |
|  | 3.708159... | 10  | 4.464464...    | 0.621093   | 3.709994... | 0.482376    |
|   |             | 15  | 4.367956...    | 3.545369   | 3.708191... | 3.738137    |
|   |             | 20  | 4.324182...    | 14.906281  | 3.708163... | 20.592531   |
|  | 4.934802... | 4   | 7.325401...    | 0.124392   | 5.612716... | 0.077315    |
|   |             | 8   | 6.618263...    | 7.222441   | 4.976796... | 7.178571    |
|   |             | 12  | 6.303035...    | 696.886298 | 4.937648... | 1105.619231 |

Table 14.1: The optimal values of (14.13) with and without Stokes constraints for Examples 14.12 and 14.13. The column “ $\max P_d$ ” displays the optimal value, whereas the column “time” gives the time, in seconds, for running the command `SemidefiniteOptimization` in `Mathematica`.

As Table 14.1 shows, the convergence with Stokes constraints is much faster than without them. The intuition is that now, with the (dual) Stokes constraints added to  $P^*$ , the function we approximate is not just the indicator function of  $K$ . A detailed explanation, for a variant of the Stokes constraints, is given in [175]. The authors prove that, when adding this new type of constraints, the optimal solution of the new  $P^*$  becomes a minimum. This eliminates any kind of Gibbs effect, and guarantees faster convergence. In [175], the authors mention that, from numerical experiments, it is reasonable to expect that the original Stokes constraints and the new Stokes constraints are equivalent, but there is no formal proof of this statement yet. We close with the remark that general semialgebraic sets fit into this framework; see [90, 175, 176].



## **Chapter 15**

### **Sampling**

This chapter is about methods for sampling from a real algebraic variety  $X \subset \mathbb{R}^n$ . In other words, we want to compute a finite subset  $S \subset X$ , which we call a *sample*. An example of a sample on a curve is shown in Figure 15.1. The sample  $S$  yields a discrete approximation of  $X$  that can be used to explore properties of the variety. For instance, if  $g : X \rightarrow \mathbb{R}$  is a function, we can find a lower bound for the optimization problem  $\max_{\mathbf{x} \in X} g(\mathbf{x})$  by computing  $\max_{\mathbf{s} \in S} g(\mathbf{s})$ , or we can estimate the integral  $\int_{\mathbf{x} \in X} g(\mathbf{x}) d\mathbf{x} / \text{vol}(X)$  (provided  $X$  has finite volume) by  $\frac{1}{|S|} \sum_{\mathbf{s} \in S} g(\mathbf{s})$ . Asmussen and Glynn [9] underline the importance of sampling methods as follows:

“Sampling-based computational methods are a fundamental part of the numerical toolset across an enormous number of different applied domains”.

In Section 15.1, we discuss in detail how samples can be used to compute topological information of  $X$ .

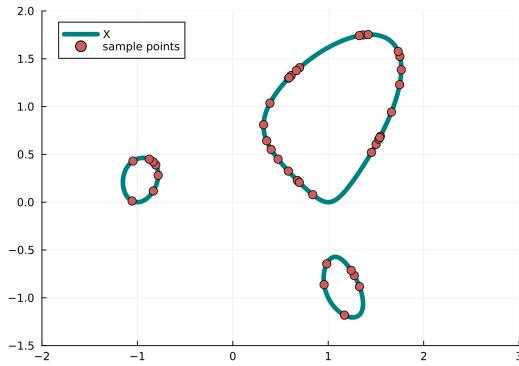


Fig. 15.1: A sample of points on the curve  $X = \{x^4 + y^4 - 2x^2 - 2xy^2 - y + 1 = 0\}$ .

In Sections 15.2 and 15.3, we discuss two approaches to sampling. The first approach concerns sampling methods with density guarantees. These methods have in common that they generate a (possibly random) sample  $S \subset X$  that is  $\varepsilon$ -dense in  $X$ .

**Definition 15.1** A finite subset  $S \subset X$  is called  $\varepsilon$ -dense in  $X$  or an  $\varepsilon$ -dense sample for  $X$  if, for all  $\mathbf{x} \in X$ , there exists  $\mathbf{s} \in S$  with  $\|\mathbf{x} - \mathbf{s}\| < \varepsilon$ .

The second approach to sampling is to generate points from a prescribed probability distribution on  $X$ . For instance, if  $X$  is compact, we could be interested in sampling from the uniform distribution.

Since we are computing random samples, it is enough to replace the variety  $X$  by its smooth locus  $X^{\text{sm}}$  and sample from  $X^{\text{sm}}$ . To keep the notation simple, though, we assume throughout the remainder of this chapter that  $X$  is a smooth manifold embedded in  $\mathbb{R}^n$ .

## 15.1 Computing the Homology from Finite Samples

We first recall a theorem due to Niyogi, Smale and Weinberger [145]. Their result gives conditions for when the homology of a smooth manifold  $X$  embedded in  $\mathbb{R}^n$  can be computed from a finite sample  $S$ . The idea is to compute the homology of the union of  $\varepsilon$ -balls  $U = \bigcup_{\mathbf{s} \in S} B_\varepsilon(\mathbf{s})$  of an  $\varepsilon$ -sample  $S$ . The homology groups of a union of balls  $U$  can be computed from the associated Čech-complex. The theorem explains how small  $\varepsilon$  must be for this to work.

**Theorem 15.2** Let  $\varepsilon > 0$ ,  $S \subset X$  be an  $\varepsilon$ -sample for  $X$ , and  $U = \bigcup_{s \in S} B_\varepsilon(s)$  be the union of  $\varepsilon$ -balls around points in  $S$ . Let  $\tau(X)$  be the reach of  $X$ . If  $\varepsilon < \sqrt{\frac{3}{20}}\tau$ , then  $X$  is a deformation retract of  $U$ . In particular, the homology of  $X$  equals the homology of  $U$ .

**Proof** See [145, Proposition 3.1].  $\square$

It was shown in [158] that we can replace the reach  $\tau(X)$  in Theorem 15.2 by what is called the *local reach*  $\tau(\mathbf{x})$ . For  $\mathbf{x} \in X$ , the local reach is defined as

$$\tau(\mathbf{x}) := \sup \{r \geq 0 \mid \text{for all } \mathbf{u} \in B_r(\mathbf{x}) \text{ there is a unique minimizer of } X \rightarrow \mathbb{R}, \mathbf{x}' \mapsto \|\mathbf{x}' - \mathbf{u}\|\}.$$

In other words, the local reach  $\tau(\mathbf{x})$  is the distance from the point  $\mathbf{x} \in X$  to the medial axis  $\text{Med}(X)$  of  $X$  (cf. Chapter 7). Hence, the global reach  $\tau(X)$  is upper bounded by the local reach at each point; in fact,  $\tau(X) = \inf_{\mathbf{x} \in X} \tau(\mathbf{x})$ . Thus, using the local reach instead of the global reach for upper bounding  $\varepsilon$  can make a significant difference. The result from [158] is: If  $\varepsilon < \frac{4}{5}\tau(s)$  for all points  $s$  in a  $\varepsilon$ -dense sample  $S \subset X$ , then  $X$  is a deformation retract of  $U = \bigcup_{s \in S} B_\varepsilon(s)$ .

Interestingly, if we only want to compute homology for small dimensions, we can relax the conditions on  $\varepsilon$  even more. For that, we recall the definition of *weak feature size* from [48]. We first need to introduce the notion of *k-bottlenecks*.

**Definition 15.3** Let  $k \geq 2$ . For  $B = \{\mathbf{x}_1, \dots, \mathbf{x}_k\} \subset X$  (with  $|B| = k$ ), let  $\Gamma(B)$  be the union of the centers of all  $(n-1)$ -spheres passing through  $\mathbf{x}_1, \dots, \mathbf{x}_k$ . If  $N_{\mathbf{x}_1}X \cap \dots \cap N_{\mathbf{x}_k}X \cap \Gamma(B) \cap \text{conv}(B) \neq \emptyset$ , we call  $B$  a *k-bottleneck* of  $X$ . Its *width* is  $\ell(B) := \inf_{\mathbf{u} \in \Gamma(B)} \|\mathbf{x}_1 - \mathbf{u}\|$ .

For instance, the 2-bottlenecks of  $X$  are precisely the bottlenecks studied in Section 7.1. The *weak feature size* is the smallest width of a  $k$ -bottleneck:

$$\text{wfs}(X) := \min_{2 \leq k \leq \text{EDdegree}(X)} \inf \{\ell(B) \mid B \text{ is a } k\text{-bottleneck of } X\}.$$

The weak feature size is always greater or equal than the reach, because the exponential map  $\varphi_\varepsilon$  in (6.10) cannot be injective when  $\varepsilon > \ell(B)$  for a  $k$ -bottleneck  $B = \{\mathbf{x}_1, \dots, \mathbf{x}_k\}$ . In fact, in that case, the point  $\mathbf{u} \in \Gamma(B)$  with  $\ell(B) = \|\mathbf{x}_1 - \mathbf{u}\|$  has  $k$  points  $(\mathbf{x}_1, \mathbf{v}_1), \dots, (\mathbf{x}_k, \mathbf{v}_k)$  in its fiber under  $\varphi_\varepsilon$ . The following was proved in [158].

**Theorem 15.4** Let  $\varepsilon < \text{wfs}(X)$  and  $S \subset X$  be an  $\varepsilon$ -dense sample. Construct the two-dimensional Vietoris-Rips complex  $C$  with vertices  $S$  as follows: Add the edge spanned by  $\mathbf{x}, \mathbf{y} \in S$  to  $C$  if and only if one of the following two conditions is satisfied:

1.  $\|\mathbf{x} - \mathbf{y}\| \leq 2\varepsilon$ , or
2.  $\|\mathbf{x} - \mathbf{y}\| \leq \sqrt{8}\varepsilon$  and there is  $\mathbf{z} \in S$  with  $\|\mathbf{x} - \mathbf{z}\| \leq 2\varepsilon$  and  $\|\mathbf{y} - \mathbf{z}\| \leq 2\varepsilon$ .

Moreover, add the 2-simplex spanned by  $\mathbf{x}, \mathbf{y}, \mathbf{z} \in S$  to  $C$  if and only if there is an edge between  $\mathbf{x}$  and  $\mathbf{y}$ , between  $\mathbf{x}$  and  $\mathbf{z}$ , and between  $\mathbf{y}$  and  $\mathbf{z}$ . Then,  $H_0(X) \cong H_0(C)$  and  $H_1(X) \cong H_1(C)$ .

**Remark 15.5** Theorem 15.4 is based on [48, Theorem 1], where it is shown that for  $\varepsilon < \text{wfs}(X)$  the tubular neighborhood  $\bigcup_{\mathbf{x} \in X} B_\varepsilon(\mathbf{x})$  is homotopy equivalent to  $X$ .

## 15.2 Sampling with Density Guarantees

This section presents two sampling algorithms that guarantee to compute an  $\varepsilon$ -sample. Both algorithms compute a sample of  $X \cap R$ , where

$$R = [a_1, b_1] \times \cdots \times [a_n, b_n] \subset \mathbb{R}^n$$

is a box. If  $X$  is compact, we can compute a box  $R$  such that  $X \subset R$  as follows. We first sample a point  $\mathbf{u} \in \mathbb{R}^n$  at random. Then, we compute the ED critical points on  $X$  with respect to  $\mathbf{u}$ ; i.e., the critical points of the Euclidean distance function  $X \rightarrow \mathbb{R}, \mathbf{x} \mapsto \|\mathbf{x} - \mathbf{u}\|$ . From this, we infer  $r := \max_{\mathbf{x} \in X} \|\mathbf{x} - \mathbf{u}\|$  and set  $R$  to be the box with center  $\mathbf{u}$  and side length  $2r$ .

The first sampling algorithm we present is from [67]. Let  $R = [a_1, b_1] \times \cdots \times [a_n, b_n]$  be a box. The basic idea for sampling from  $X \cap R$  is to sample points  $\mathbf{u} \in \mathbb{R}^n$  and then to collect the ED critical points with respect to  $\mathbf{u}$ . The complexity of this approach therefore depends on the Euclidean Distance Degree of  $X$ .

The algorithm in [67] works recursively by dividing the sides of the box  $R$  in half, thus splitting  $R$  into  $2^n$  subboxes. In addition, one implements a database  $\mathcal{D}$  that contains information about all the regions in  $R$  that already have been covered by at least one sample point. Algorithm 4 provides a complete description. The correctness of that algorithm is proved in [67].

**Theorem 15.6 (Theorem 4.4 in [67])** *Algorithm 4 terminates and outputs an  $\varepsilon$ -dense sample of  $X \cap R$ .*

---

**Algorithm 4:** The algorithm from [67].

---

```

1 Input: A real algebraic variety  $X \subset \mathbb{R}^n$ , a real number  $\varepsilon > 0$ , and a box  $R = [a_1, b_1] \times \cdots \times [a_n, b_n]$ 
2 Output: An  $\varepsilon$ -dense sample  $S \subset X \cap R$ .
3 Initialize  $S = \emptyset$  and  $\mathcal{D} = \emptyset$ . The set  $S$  will contain the sample points. The set  $\mathcal{D}$  serves as a database containing
   balls in  $\mathbb{R}^n$  that have already been covered in the process of the algorithm.
4 for each subbox  $R'$  of  $R$  that is not yet covered do
5   Compute the midpoint  $\mathbf{u}$  of  $R'$ .
6   Compute the real ED critical points  $E \subset X$  with respect to  $\mathbf{u}$  and compute  $r := \min_{\mathbf{x} \in X} \|\mathbf{x} - \mathbf{u}\|$ .
7   Add the points in  $E$  to  $S$ .
8   Add  $B_r(\mathbf{u})$  to  $\mathcal{D}$  (this open ball does not contain any point in  $X$ , so we do not need to consider this region any
      further and can label it as being covered).
9   Add  $B_\varepsilon(\mathbf{y})$  for  $\mathbf{y} \in E$  to  $\mathcal{D}$ .
10  if the union of balls in  $\mathcal{D}$  cover  $R'$  then
11    | Label  $R'$  and all of its subboxes as covered.
12  else
13    | Split  $R'$  into  $2^n$  smaller subboxes.
14  end
15 end
16 if all subboxes of  $R$  are labeled as covered then
17  | return  $S$ .

```

---

The second algorithm we present is from [158]. That algorithm is also based on computing ED critical points on  $X$ , but in addition adds linear slices to the sampling. We need a few definitions. We set  $d := \dim X$ . For every  $1 \leq k \leq d$ , denote by  $\mathcal{T}_k$  the set of subsets of  $\{1, \dots, n\}$  with  $k$  elements. Given a set  $T = \{t_1, \dots, t_k\} \in \mathcal{T}_k$ , we let  $V_T \subseteq \mathbb{R}^n$  be the  $k$ -dimensional coordinate plane spanned by  $\mathbf{e}_{t_1}, \dots, \mathbf{e}_{t_k}$ . For  $\delta > 0$ , consider the grid

$$G_T(\delta) := \{\delta \cdot (a_1 \cdot \mathbf{e}_{t_1} + \cdots + a_d \cdot \mathbf{e}_{t_k}) \mid a_1, \dots, a_k \in \mathbb{Z}\} \cong \delta \cdot \mathbb{Z}^k.$$

Let  $\pi_T : \mathbb{R}^n \rightarrow V_T$  be the projection. Then, the linear spaces  $\pi_T^{-1}(g)$  for  $g \in G_T(\delta)$  are given by the faces of a cubical tessellation with side length  $\delta$ .

Let  $B(X)$  be the width of the smallest bottleneck of  $X$ . The algorithm from [158] takes as input a number  $0 < \delta < \frac{1}{\sqrt{n}} \min\{\varepsilon, 2B(X)\}$ . Then, the sample is given by

$$S_\delta := \bigcup_{T \in \mathcal{T}_d} \bigcup_{g \in G_T(\delta)} X \cap \pi_T^{-1}(g); \quad (15.1)$$

i.e.,  $S_\delta$  consists of the points that are obtained by intersecting  $X$  with the collection of linear spaces  $\pi_T^{-1}(g)$  ranging over  $T \in \mathcal{T}_d$  and  $g \in G_T(\delta)$ . The dimension of  $\pi_T^{-1}(g)$  equals the codimension of  $X$ . To ensure transversal intersections, we can always modify  $G_T(\delta)$  by a random translation.

If  $d = \dim X > 1$ , the algorithm computes an additional sample. First, we sample a random point  $\mathbf{u} \in \mathbb{R}^n$ . Denote by  $E(T, g, \mathbf{u})$  the ED critical points on  $X \cap \pi_T^{-1}(g)$  with respect to  $\mathbf{u}$ . The additional sample is

$$S'_\delta = \bigcup_{k=1}^{d-1} \bigcup_{T \in \mathcal{T}_k} \bigcup_{g \in G_T(\delta)} E(T, g, \mathbf{u}). \quad (15.2)$$

The motivation for this extra sample is that  $E(T, g, \mathbf{u})$  contains a point on every connected component of  $X \cap \pi_T^{-1}(g)$ . The algorithm is summarized in Algorithm 5.

**Theorem 15.7 (Theorem 4.6 in [158])** Algorithm 5 outputs an  $\varepsilon$ -dense sample of  $X \cap R$ .

---

**Algorithm 5:** The algorithm from [158].

---

```

1 Input: A real algebraic variety  $X \subset \mathbb{R}^n$ , a real number  $\delta > 0$  with  $\delta < \frac{1}{\sqrt{n}} \min\{\varepsilon, 2B(X)\}$ , and a box
    $R = [a_1, b_1] \times \cdots \times [a_n, b_n]$ .
2 Output: An  $\varepsilon$ -dense sample  $S \subset X \cap R$ .
3 Initialize  $S_\delta = \emptyset$  and  $S'_\delta = \emptyset$ .
4 Set  $d := \dim X$ .
5 for  $T \in \mathcal{T}_d$  and  $g \in G_T(\delta)$  do
6   Compute  $X \cap \pi_T^{-1}(g)$ .
7   Add the points in  $X \cap \pi_T^{-1}(g)$  to  $S_\delta$ .
8 end
9 if  $d > 1$  then
10  for  $1 \leq k < d$  do
11    for  $T \in \mathcal{T}_k$  and  $g \in G_T(\delta)$  do
12      Sample a random point  $\mathbf{u} \in \mathbb{R}^n$ .
13      Compute  $E(T, g, \mathbf{u})$ , which are the ED critical points on  $X \cap \pi_T^{-1}(g)$  with respect to  $\mathbf{u}$ .
14      Add the points in  $E(T, g, \mathbf{u})$  to  $S'_\delta$ .
15    end
16  end
17 Set  $S := S_\delta \cup S'_\delta$ .
18 return  $S$ .

```

---

## 15.3 Sampling from Probability Distributions

In this section, we discuss algorithms for sampling from a probability measure  $\pi$  on  $X$ . In this context,  $\pi$  is called the *target distribution*. We will review one popular class of methods for sampling from probability distributions on  $X$ , namely *Markov-Chain Monte Carlo* (MCMC) methods. In the following,  $\mathcal{A}$  denotes the  $\sigma$ -algebra of  $X$ . Think of  $\mathcal{A}$  as the collection of subsets of  $X$  that we can assign a measure to; the elements in  $\mathcal{A}$  are called *measurable*. If  $\mu : \mathcal{A} \rightarrow \mathbb{R}$  is a measure and  $f : X \rightarrow \mathbb{R}$  is a measurable function, we can integrate  $f$  against  $\mu$ . This integral is denoted  $\int_A f(\mathbf{x}) \mu(d\mathbf{x})$  for  $A \in \mathcal{A}$ . The Lebesgue measure on  $X$  is simply denoted  $d\mathbf{x}$ . If  $\mu$  is a probability measure with a probability density  $\phi$ , then  $\mu(d\mathbf{x}) = \phi(\mathbf{x})d\mathbf{x}$ .

**Example 15.8** For a measurable set  $A \in \mathcal{A}$ , let  $\text{vol}(A) = \int_A d\mathbf{x}$  be the volume of  $A$ . Assume that  $X$  is compact. Then  $\text{vol}(A) < \infty$  and the probability distribution with probability measure  $\pi(A) := \text{vol}(A)/\text{vol}(X)$  is called the *uniform distribution* on  $X$ . For instance, the sample in Figure 15.1 consists of i.i.d. (independent and identically distributed) points from the uniform distribution on the plane curve  $X = \{x^4 + y^4 - 2x^2 - 2xy^2 - y + 1 = 0\}$ .

Before we discuss sampling algorithms, let us briefly recall the following result by Niyogi, Smale and Weinberger [145], which implies that sampling from the uniform distribution on  $X$  can be used to generate  $\varepsilon$ -samples with high probability.

**Proposition 15.9** Let  $\varepsilon > 0$  and  $0 < \delta < 1$ . Suppose  $X$  is compact of dimension  $d := \dim X$  and with positive reach  $\tau(X) > 0$ . Set

$$\theta := \arcsin\left(\frac{\varepsilon}{\tau}\right), \quad \omega := \frac{\text{vol}(X)}{\cos^d \theta}, \quad \beta_1 := \frac{\omega}{\text{vol}(B_\varepsilon(\mathbf{0}))}, \quad \beta_2 := \frac{\omega}{\text{vol}(B_{\varepsilon/8}(\mathbf{0}))}$$

(here,  $B_\varepsilon(\mathbf{0})$  is the unit ball of radius  $\varepsilon$ ). Let  $n > \beta_1(\log \beta_2 - \log \delta)$ . For a sample of i.i.d. points  $S = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  from the uniform distribution on  $X$ , we have  $\text{Prob}(S \text{ is an } \varepsilon\text{-sample for } X) > 1 - \delta$ .

Let us now put the focus on sampling by using MCMC methods. These methods set up a *Markov process* on  $X$ . Let us recall some basic definitions from the theory of Markov chains; see, e.g., [133, Chapter 3]. In this context,  $X$  is also called a *state space*. A *Markov kernel* is a map  $p : X \times \mathcal{A} \rightarrow [0, 1]$ , such that

1.  $p(\mathbf{x}, \cdot)$  is a probability measure for all  $\mathbf{x} \in X$ ;
2.  $p(\cdot, A)$  is a measurable function for all  $A \in \mathcal{A}$ .

A stochastic process  $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \dots$  on  $X$  is a sequence of random points on  $X$ . Markov processes encompass a special type of stochastic processes. Let  $\mathbf{x} \in X$  be a fixed point. A (time-homogeneous) Markov process with starting point  $\mathbf{x}$  is the stochastic process defined by  $\mathbf{x}_0 = \mathbf{x}$ , and for  $k \geq 1$  by the following iterated integral:

$$\text{Prob}(\mathbf{x}_k \in A_k, \dots, \mathbf{x}_1 \in A_1 \mid \mathbf{x}_0 = \mathbf{x}) := \int_{\mathbf{y}_1 \in A_1} \dots \int_{\mathbf{y}_{k-1} \in A_{k-1}} p(\mathbf{x}, d\mathbf{y}_1) p(\mathbf{y}_1, d\mathbf{y}_2) \dots p(\mathbf{y}_{k-1}, A_k).$$

For fixed  $\mathbf{z} \in X$ , we have

$$\text{Prob}(\mathbf{x}_k \in A \mid \mathbf{x}_{k-1} = \mathbf{z}, \mathbf{x}_{k-2}, \dots, \mathbf{x}_0) = \text{Prob}(\mathbf{x}_k \in A \mid \mathbf{x}_{k-1} = \mathbf{z}) = p(\mathbf{z}, A); \quad (15.3)$$

see, e.g., [133, Proposition 3.4.3]. Equation (15.3) is called the *Markov property*. It means that the probability law of the next point in the process only depends on the current state  $\mathbf{z}$ , but not on earlier states. Due to its role in (15.3), the kernel  $p$  is also called *transition probability*.

**Example 15.10 (Markov Chains in  $\mathbb{R}^2$ )** Let us consider two examples of Markov Chains in  $\mathbb{R}^2$  starting at the origin. For the first, we take the kernel  $p_1(\mathbf{x}, d\mathbf{y}) = (2\pi)^{-1} \cdot \exp(-\frac{1}{2}\|\mathbf{y}\|^2) d\mathbf{y}$ . The stochastic process arising from this passes from the state  $\mathbf{x} \in \mathbb{R}^2$  to the next state by sampling a normal vector in  $\mathbb{R}^2$  with covariance matrix the identity and mean value  $\mathbf{0}$ . In particular, the transition probability is independent of the current state.

The second example has the kernel  $p_2(\mathbf{x}, d\mathbf{y}) = (2\pi)^{-1} \exp(-\frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2) d\mathbf{y}$ . In this case, passing from a state  $\mathbf{x} \in \mathbb{R}^2$  to the next state works by sampling a normal vector in  $\mathbb{R}^2$  with covariance matrix the identity and mean value  $\mathbf{x}$ . We can simulate the first  $n = 10$  steps in this chain in Julia as follows.

```
x0 = [0; 0]
states = []
push!(states, x0)
```

```

n = 10
for k in 1:n
    x = states[k]
    y = x + randn(2)
    push!(states, y)
end

```

Figure 15.2 shows one realization of this chain.

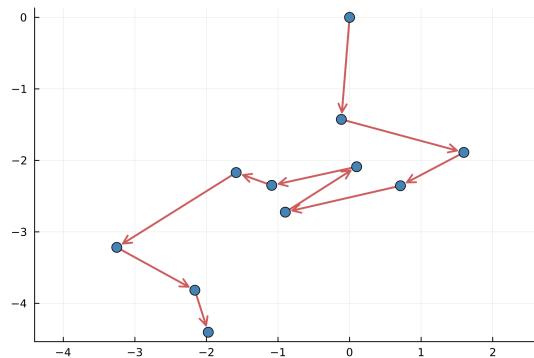


Fig. 15.2: The picture shows one sample of the first 10 steps of the Markov Chain with transition probability at state  $\mathbf{x}$  given by sampling a normal vector in  $\mathbb{R}^2$  with covariance matrix the identity and mean value  $\mathbf{x}$ . See Example 15.10.

**Example 15.11 (A Markov Chain on a Real Variety)** Let  $X \subset \mathbb{R}^n$  be a real variety of dimension  $d$  and degree at least two. We can create a Markov chain on  $X$  as follows. Suppose that chain is in state  $\mathbf{x} \in X$ . We sample a random linear space  $L$  of complementary dimension passing through  $\mathbf{x}$ . Then, we sample a point uniformly at random from  $X \cap L$ . This is the next state. As an example, we take the surface defined by  $z - xy = 0$ . Given a point  $\mathbf{x}$  on this surface, we sample the line  $L = \{A\mathbf{x} = \mathbf{b}\}$  by sampling  $A$  with Gaussian entries and setting  $\mathbf{b} = A\mathbf{x}$ . In addition, we only accept states in the box  $R = [-8, 8] \times [-8, 8] \times [-64, 64]$ . The first few steps in this chain starting at  $\mathbf{x}_0 = (0, 0, 0)$  are implemented in Julia [21] as follows.

```

using HomotopyContinuation
@var x y z
f = System([z - x * y], variables = [x; y; z])
is_in_R(p) = abs(p[1]) < 8 && abs(p[2]) < 8 && abs(p[3]) < 64

n = 10
x0 = [0.0; 0.0; 0.0]
states = [x0]
for k in 1:n
    p = last(states)
    A = randn(2,3); b = A*p
    L = LinearSubspace(A, b)

    S = solve(f, target_subspace = L)
    points = real_solutions(S)
    filter!(is_in_R, points)
    push!(states, rand(points))
end

```

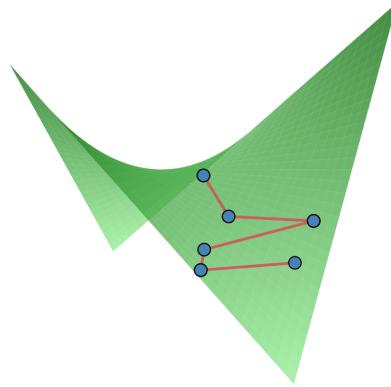


Fig. 15.3: The picture shows first few steps of the Markov Chain from Example 15.11 on the surface  $z - xy = 0$ .

One realization of this example is shown in Figure 15.3.

Let us now work towards sampling. We will recall results from the survey [157] by Roberts and Rosenthal. A probability measure  $\pi$  on  $X$  is called a *stationary distribution* if

$$\int_{\mathbf{x} \in X} p(\mathbf{x}, A) \pi(d\mathbf{x}) = \pi(A) \quad \text{for all } A \in \mathcal{A}. \quad (15.4)$$

The idea of MCMC methods for sampling from a probability distribution  $\pi$  is to set up a Markov process on  $X$  starting at  $\mathbf{x}$  such that  $\pi$  is stationary and such that the probability measure

$$\mu^k(\mathbf{x}, \cdot) : A \mapsto \text{Prob}(\mathbf{x}_k \in A \mid \mathbf{x}_0 = \mathbf{x})$$

converges to  $\pi$  as  $k \rightarrow \infty$ . Convergence is measured by the *total variation distance*. The total variation distance between two measures  $\mu$  and  $\nu$  is  $d_{\text{TV}}(\mu, \nu) := \sup_{A \in \mathcal{A}} |\mu(A) - \nu(A)|$ . Thus, we want to define a Markov process starting at a point  $\mathbf{x} \in X$  such that  $\lim_{k \rightarrow \infty} d_{\text{TV}}(\pi, \mu^k(\mathbf{x}, \cdot)) = 0$ . The key properties for achieving this are *irreducibility* and *aperiodicity*.

**Definition 15.12** A Markov chain with kernel  $p$  is called *irreducible* if, for all states  $\mathbf{x} \in X$  and all measurable sets  $A \in \mathcal{A}$  with  $\text{vol}(A) > 0$ , there exists  $k \in \mathbb{N}$  such that  $\mu^k(\mathbf{x}, A) > 0$ .

The interpretation of irreducibility is that all sets with positive volume can eventually be reached by the Markov chain starting at any point  $\mathbf{x} \in X$ .

*Remark 15.13* One can replace the Lebesgue measure in the definition by any other measure  $\phi$ . In that case, one speaks of  $\phi$ -irreducible Markov chains.

**Definition 15.14** Consider a Markov chain with kernel  $p$  and suppose that it has a stationary distribution  $\pi$ . Let  $d \geq 2$ . We call the chain *periodic* with period  $d$  if there exist disjoint subsets  $A_1, \dots, A_d \in \mathcal{A}$  that satisfy

1.  $\pi(A_i) > 0$  and
2.  $p(\mathbf{x}, A_{i+1} \bmod d) = 1$  for all  $\mathbf{x} \in A_i$

for all  $i$ . Otherwise, the chain is called *aperiodic*.

The geometric interpretation of aperiodicity is that the Markov process does not move periodically between disjoint subset  $A_1, \dots, A_d$ . Given an irreducible and aperiodic Markov chain with stationary distribution  $\pi$ , we have the following convergence result.

**Theorem 15.15** *Let  $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \dots$  be a Markov chain with kernel  $p$  and stationary distribution  $\pi$ . Suppose that the chain is irreducible and aperiodic. Then,*

$$\lim_{k \rightarrow \infty} d_{\text{TV}}(\pi, \mu^k(\mathbf{x}, \cdot)) = 0$$

for almost all  $\mathbf{x} \in X$ .

**Proof** See [157, Theorem 4]. □

**Remark 15.16** For this theorem, it is enough that  $X$  is a state space with countably generated  $\sigma$ -algebra. Smooth submanifolds of  $\mathbb{R}^n$  always satisfy this hypothesis.

The next result gives, under some assumptions, the speed of convergence in Theorem 15.15. Roberts and Rosenthal [157] attribute this theorem to Doeblin, Doob and also to Markov.

**Theorem 15.17** *Consider a Markov Chain with stationary distribution  $\pi$ . Suppose that there exist  $n \in \mathbb{N}$ ,  $\varepsilon > 0$  and a probability measure  $\nu$  on  $X$  such that*

$$\mu^n(\mathbf{x}, A) \geq \varepsilon \cdot \nu(A) \quad \text{for all } \mathbf{x} \in X \text{ and } A \in \mathcal{A}.$$

*Then, we have*

$$d_{\text{TV}}(\pi, \mu^k(\mathbf{x}, \cdot)) < (1 - \varepsilon)^{\lfloor \frac{k}{n} \rfloor}.$$

**Proof** See [157, Theorem 8]. □

The condition in Theorem 15.17 intuitively means that all  $n$  step transitions have a component of probability at least  $\varepsilon$  in common.

Theorems 15.15 and 15.17 have the following algorithmic consequence. For sampling from a distribution  $\pi$ , we set up an irreducible and aperiodic Markov chain with stationary distribution  $\pi$ . Then, if we let the Markov chain run long enough, the points in the process will have a probability distribution close to  $\pi$ . This is the idea underlying MCMC methods. The key task is thus to find and implement such a chain. The first subtask is to find a chain whose stationary distribution is  $\pi$ . It is often easier to show that a Markov chain is *reversible* with respect to  $\pi$ .

**Definition 15.18** A Markov chain with kernel  $p$  is reversible with respect to a probability distribution  $\pi$  if  $\int_{\mathbf{x} \in B} p(\mathbf{x}, A) \pi(d\mathbf{x}) = \int_{\mathbf{x} \in A} p(\mathbf{x}, B) \pi(d\mathbf{x})$  for all  $A, B \in \mathcal{A}$ .

**Lemma 15.19** *If a Markov chain with kernel  $p$  is reversible with respect to a probability distribution  $\pi$ , then  $\pi$  is a stationary distribution.*

**Proof** We have to check the equation in (15.4). Fix  $A \in \mathcal{A}$ . Then

$$\int_{\mathbf{x} \in X} p(\mathbf{x}, A) \pi(d\mathbf{x}) = \int_{\mathbf{x} \in A} p(\mathbf{x}, X) \pi(d\mathbf{x}) = \pi(A),$$

since  $p(\mathbf{x}, \cdot)$  is a probability measure, so that  $p(\mathbf{x}, X) = 1$  for all  $\mathbf{x} \in X$ . □

Using Lemma 15.19, it is straightforward to show that the *Metropolis Hastings algorithm* (Algorithm 6) creates a Markov chain with stationary distribution  $\pi$ ; see, e.g., [157, Proposition 2].

The Metropolis Hastings algorithm works for a target distribution  $\pi$  that has a density  $\phi$ . The basic idea is to take another Markov chain whose kernel  $p(\mathbf{x}, A)$  has a density  $q(\mathbf{x}, \mathbf{y})$ ; i.e.,  $p(\mathbf{x}, d\mathbf{y}) = q(\mathbf{x}, \mathbf{y}) d\mathbf{y}$ . The density  $q$  is called a *proposal density*. Sampling from the proposal density creates a random proposal point  $\mathbf{y}$ , which is either accepted or rejected depending on how likely it is that the proposal point  $\mathbf{y}$  was sampled from  $\pi$ . Notice that in the algorithm we only need to evaluate  $\phi(\mathbf{y})$  and  $q(\mathbf{x}, \mathbf{y})$  up to scaling.

---

**Algorithm 6:** The Metropolis Hastings algorithm.

---

```

1 Input: A probability measure  $\pi$  on  $X$  with a density  $\phi(\mathbf{y})$ . A Markov kernel  $p(\mathbf{x}, A)$  on  $X$  with density  $q(\mathbf{x}, \mathbf{y})$ . A fixed starting point  $\mathbf{x} \in X$ .
2 Output: A Markov chain on  $X$  with stationary distribution  $\pi$ .
3 Set  $\mathbf{x}_0 = \mathbf{x}$ .
4 for  $k = 0, 1, 2, \dots$  do
5   Sample  $\mathbf{y} \sim p(\mathbf{x}_k, \cdot)$ .
6   if  $\phi(\mathbf{x}_k) = 0$  or  $q(\mathbf{x}_k, \mathbf{y}) = 0$  then
7     Set  $w(\mathbf{x}_k, \mathbf{y}) = 0$ 
8   else
9     Compute  $w(\mathbf{x}_k, \mathbf{y}) = \min \left\{ 1, \frac{\phi(\mathbf{y}) \cdot q(\mathbf{y}, \mathbf{x}_k)}{\phi(\mathbf{x}_k) \cdot q(\mathbf{x}_k, \mathbf{y})} \right\}$ .
10  end
11  Sample a Bernoulli random variable  $\beta \in \{0, 1\}$  with  $\text{Prob}\{\beta = 1\} = w(\mathbf{x}_k, \mathbf{y})$ .
12  if  $\beta = 1$  then
13    Set  $\mathbf{x}_{k+1} := \mathbf{y}$ .
14  else
15    Return to line 5.
16  end
17 end

```

---

**Example 15.20 (The Symmetric Metropolis Algorithm)** For a symmetric density  $q(\mathbf{x}, \mathbf{y}) = q(\mathbf{y}, \mathbf{x})$ , the Metropolis Hastings algorithm (Algorithm 6) is called *Symmetric Metropolis Algorithm*. For instance, the Markov kernel with density  $q(\mathbf{x}, \mathbf{y}) \propto \exp(-\frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2)$  from Example 15.10 is symmetric.

**Example 15.21 (Random Walks)** We speak of a *Random Walk Metropolis Hastings Algorithm* if the proposal density  $q(\mathbf{x}, \mathbf{y})$  has the form  $q(\mathbf{x} - \mathbf{y})$ . An example for this is the density  $q(\mathbf{x}, \mathbf{y}) \propto \exp(-\frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2)$  from Example 15.10.

**Example 15.22 (Independence Sampler)** We call Algorithm 6 an *independence sampler* if  $q(\mathbf{x}, \mathbf{y})$  does not depend on  $\mathbf{x}$ . In this case, the samples obtained from Algorithm 6 are independent.

**Example 15.23 (The Langevin Algorithm)** The Langevin Algorithm works for sampling in  $X = \mathbb{R}^n$ . Suppose that the density  $\phi(\mathbf{y})$  of the target distribution  $\pi$  in Algorithm 6 is differentiable. Denote by  $\nabla\phi(\mathbf{y})$  the gradient. The Langevin Algorithm generates a proposal  $\mathbf{y}$  by sampling  $\mathbf{y} \sim N(\mathbf{x}_k + \delta \cdot \nabla\phi(\mathbf{x}_k), 2\delta)$ , where  $\mathbf{x}_k$  is the current state. This choice is motivated by a discrete approximation to a Langevin diffusion processes.

**Example 15.24 (Sampling points that are close to a variety)** This example is inspired by an idea first suggested by Jon Hauenstein and David Kahle [86]. We can use Algorithm 6 for sampling points near a variety as follows. Choose a box  $R \subset \mathbb{R}^n$  and consider the probability measure  $\pi$  on  $R$  given by the density

$$\phi(\mathbf{u}) \propto \exp \left( -\frac{1}{2\sigma^2} d(\mathbf{u}, X)^2 \right),$$

where  $\sigma^2 > 0$  and  $d(\mathbf{u}, X) = \min_{\mathbf{x} \in X} \|\mathbf{x} - \mathbf{u}\|$  is the Euclidean distance from  $\mathbf{u}$  to  $X$ . Sampling from  $\pi$  produces points that have a high probability of being close to  $X$ . Let us sample from this distribution for the Trott curve  $X = \{144(x^4 + y^4) - 225(x^2 + y^2) + 350x^2 * y^2 + 81 = 0\}$  and the box  $R = [-2, 2] \times [-2, 2]$ . We implement the Metropolis-Hastings algorithm in Julia [21] using the proposal density

$$q(\mathbf{x}, \mathbf{y}) \propto \exp(-\frac{1}{2} \|\mathbf{y}\|^2);$$

i.e., the proposals are sampled from the two-dimensional standard normal distribution (so we use an independence sampler as described in Example 15.22). First, we implement the system that we need to solve in order to compute  $d(\mathbf{u}, X)$ , and we solve it for a general complex point  $\mathbf{u}$  that we will use later for doing parameter homotopies.

```
using HomotopyContinuation
@var x[1:2] u[1:2] l
f = 144(x[1]^4 + x[2]^4) - 225(x[1]^2 + x[2]^2) + 350x[1]^2*x[2]^2 + 81
df = differentiate(f, x)
F = System([f; df - 1 .* (x-u)], variables = [x; 1], parameters = u)

uC = randn(ComplexF64, 2)
SC = solve(F, target_parameters = uC)
```

Next, we implement the densities  $q$  and  $\phi$  (for  $\sigma^2 = 1/100$ ) and a function that detects whether or not a point is in  $R$ .

```
is_in_R(p) = abs(p[1])<2 && abs(p[2])<2
q(u, v) = exp(-1/2 * norm(v)^2)
sigma_sq = 1/100
function phi(u)
    S = solve(F, solutions(SC), start_parameters = uC,
              target_parameters = u,
              show_progress = false)
    R = map(s -> s[1:2], real_solutions(S))
    d = minimum([norm(r - u) for r in R])
    exp(-d^2 / (2*sigma_sq))
end
```

Finally, we run the Metropolis-Hastings algorithm for 10.000 steps. The result can be seen in Figure 15.4.

```
n = 10000
u0 = randn(2)
states = [u0]
for k in 1:n
    uk = last(states)
    v = randn(2)
    if is_in_R(v)
        a = min(1, phi(v) * q(v, uk) / (phi(uk) * q(uk, v)))
        b = rand()
        if a > b
            push!(states, v)
        end
    end
end
```

One issue with Algorithm 6 when sampling from a nonlinear manifold  $X$  is to find a suitable proposal distribution with a density. Take, for instance, the Markov chain from Example 15.11, where the next step  $\mathbf{x}_{k+1} \in X$  was computed from the current step  $\mathbf{x}_k$  by taking a random linear space  $L$  through  $\mathbf{x}_k$  and sampling uniformly from the intersection points in  $X \cap L$ . It is straightforward to describe the generation of this random variable. But it is not clear how to compute its density (or if such a density even exists).

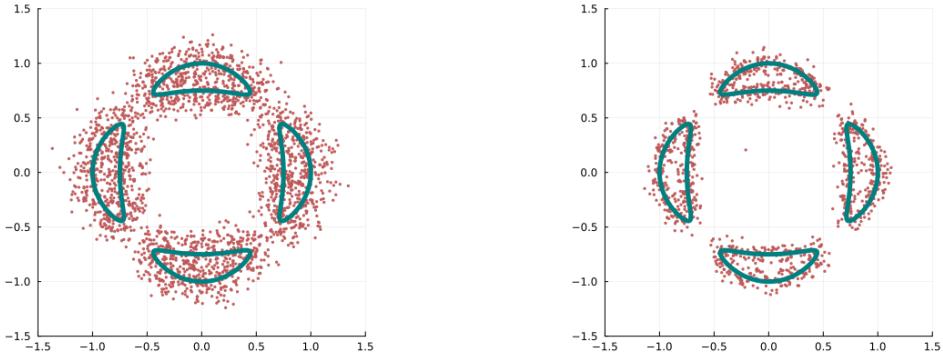


Fig. 15.4: Two samples of points near the Trott curve. They were generated using the approach from Example 15.24 with  $\sigma^2 = 1/100$  (left) and  $\sigma^2 = 1/400$  (right).

In such a scenario, we must prove that the chain is reversible with respect to  $\pi$ . A further complication arises when  $X$  is not connected. In that case, the proposals must be chosen in a way that no connected component of  $X$  is missed.

In the remainder of this section, we present two alternative approaches to Metropolis Hastings that do not require the knowledge of a proposal density. The first is the algorithm in [26] using an independence sampler. Second, we consider the algorithm in [126] based on a random walk.

The algorithm in [26] is based on the idea of linear slicing. However, instead of considering a Markov chain in  $X$ , the paper considers a Markov chain in the Grassmannian  $\text{Gr}(c, \mathbb{R}^n)$ , where  $c = \text{codim } X$ . The algorithm is an independence sampler. The following theorem describes the density with which we have to sample  $L \in \text{Gr}(c, \mathbb{R}^n)$ . As before, we assume that the target distribution  $\pi$  has a density  $\phi(\mathbf{y})$ . For  $(A, \mathbf{b}) \in \mathbb{R}^{d \times n} \times \mathbb{R}^d$ , we introduce the new function

$$\bar{\phi}(A, \mathbf{b}) := \sum_{\mathbf{x} \in X : A\mathbf{x} = \mathbf{b}} \frac{\phi(\mathbf{x})}{\alpha(\mathbf{x})}, \quad \text{where} \quad \alpha(\mathbf{x}) := \frac{\sqrt{1 + \langle \mathbf{x}, P_{\mathbf{x}} \mathbf{x} \rangle}}{(1 + \|\mathbf{x}\|^2)^{(d+1)/2}} \frac{\Gamma\left(\frac{d+1}{2}\right)}{\sqrt{\pi^{d+1}}},$$

and  $P_{\mathbf{x}}$  is the orthogonal projection onto the normal space  $N_{\mathbf{x}}X$ . As before,  $d = \dim X$  is the dimension of  $X$ . The additional factor  $\alpha$  is related to the change of variables when embedding  $\mathbb{R}^n$  into the  $n$ -dimensional real projective space.

**Theorem 15.25** Let  $d := \dim X$  and  $c := n - d = \text{codim } X$ . Let  $\varphi(A, \mathbf{b})$  be the probability density for which the entries of  $(A, \mathbf{b}) \in \mathbb{R}^{d \times n} \times \mathbb{R}^d$  are i.i.d. standard Gaussian. Denote

$$\psi(A, \mathbf{b}) := \frac{\varphi(A, \mathbf{b}) \cdot \bar{\phi}(A, \mathbf{b})}{\mathbb{E}_{(A, \mathbf{b}) \sim \varphi} \bar{\phi}(A, \mathbf{b})}.$$

Then,  $\psi$  is a probability density and the random linear space  $L := \{\mathbf{x} \in \mathbb{R}^n \mid A\mathbf{x} = \mathbf{b}\} \in \text{Gr}(c, \mathbb{R}^n)$  for  $(A, \mathbf{b}) \sim \psi$  has the properties:

1.  $X \cap L$  is finite with probability one.
2. If we choose one of the finitely many points in  $\mathbf{x} \in X \cap L$  with probability

$$\text{Prob}\{\mathbf{x}\} := \frac{\phi(\mathbf{x})}{\alpha(\mathbf{x}) \cdot \bar{\phi}(A, \mathbf{b})},$$

the random point  $\mathbf{x}$  is distributed according to the density  $\phi$ .

**Proof** See [26, Theorem 1.1].  $\square$

We review another algorithm from the literature. The algorithm in [126] is similar to the Metropolis-Hastings algorithm. The basic idea is as follows. Suppose we want to sample from a target density  $\phi(\mathbf{y})$ . Write this as  $\phi(\mathbf{y}) = \exp(-V(\mathbf{y}))$  and assume that  $V(\mathbf{y})$  is smooth. The function  $V$  is called a *potential function*. Given a point  $\mathbf{x} \in X$ , we create a proposal by sampling first in a random tangent direction  $\mathbf{v} \in T_{\mathbf{x}}X$  and then computing the intersection of  $X$  with the random linear space

$$L = \mathbf{x} + \mathbf{v} + N_{\mathbf{x}}X \in \text{Gr}(c, \mathbb{R}^n).$$

This creates a Markov chain on  $X$ , but as we have discussed before, it is not clear how to compute the density for this random proposal. Instead, the authors in [126] prove directly reversibility of their Markov chain, so that they can apply Lemma 15.19. Algorithm 7 shows a (simplified version) of their algorithm.

---

**Algorithm 7:** The algorithm from [126].

---

```

1 Input: A probability measure  $\pi$  on  $X$  with a density  $\exp(-V(\mathbf{y}))$ , where  $V$  is smooth. A variance
    parameter  $\sigma^2 > 0$ . A fixed starting point  $\mathbf{x} \in X$ .
2 Output: A Markov chain on  $X$  with stationary distribution  $\pi$ .
3 Set  $\mathbf{x}_0 = \mathbf{x}$ .
4 for  $k = 0, 1, 2, \dots$  do
5   Randomly draw  $\mathbf{v} \in T_{\mathbf{x}_k}X$  by sampling  $\mathbf{v}$  from the multivariate normal distribution on  $T_{\mathbf{x}_k}X$  with mean  $\mathbf{0}$  and
      covariance matrix  $\sigma^2 \cdot I$ .
6   Set  $L = \mathbf{x}_k + \mathbf{v} + N_{\mathbf{x}_k}X$ .
7   Sample a point  $\mathbf{y} \in X \cap L$  uniformly.
8   Compute  $\mathbf{w} \in T_{\mathbf{y}}X$  such that  $\mathbf{x}_k \in K := \mathbf{y} + \mathbf{w} + N_{\mathbf{y}}X$ .
9   Compute  $w(\mathbf{x}_k, \mathbf{y}) = \min \left\{ 1, \frac{|X \cap L|}{|X \cap K|} \cdot \exp(-(V(\mathbf{y}) - V(\mathbf{x})) \cdot \exp(-\frac{1}{2\sigma^2} (\|\mathbf{w}\|^2 - \|\mathbf{v}\|^2)) \right\}$ .
10  Sample a Bernoulli random variable  $\beta \in \{0, 1\}$  with  $\text{Prob}\{\beta = 1\} = w(\mathbf{x}_k, \mathbf{y})$ .
11  if  $\beta = 1$  then
12    | Set  $\mathbf{x}_{k+1} := \mathbf{y}$ .
13  else
14    | Return to line 5.
15  end
16 end

```

---

**Remark 15.26** The sampling of the random tangent vector  $\mathbf{v}$  in line 5 of Algorithm 7 can be achieved as follows. Let  $U \in \mathbb{R}^{n \times d}$  be a matrix whose columns form an orthonormal basis of  $T_{\mathbf{x}}X$ . Such a matrix can be computed by using the Gram-Schmidt algorithm. Sample  $\mathbf{u} \in \mathbb{R}^d$  with i.i.d.  $N(0, \sigma^2)$ -entries. Then  $\mathbf{v} = U\mathbf{u}$ .

The following theorem asserts that Algorithm 7 works correctly.

**Theorem 15.27** Algorithm 7 produces a Markov chain that is reversible with respect to  $\pi$ . In particular, by Lemma 15.19, the Markov chain has  $\pi$  as stationary distribution.

**Proof** See [126, Theorem 1].  $\square$



## References

1. H. Abo, A. Seigal, and B. Sturmfels. *Eigenconfigurations of tensors*, pages 1–25. 01 2017.
2. Daniele Agostini, Taylor Brysiewicz, Claudia Fevola, Lukas Kühne, Bernd Sturmfels, and Simon Telen. Likelihood degenerations. *Adv. Math.*, 414:Paper No. 108863, 39, 2023. With an appendix by Thomas Lam.
3. Chris Aholt and Luke Oeding. The ideal of the trifocal variety. *Mathematics of Computation*, 83, 05 2012.
4. Chris Aholt, Bernd Sturmfels, and Rekha Thomas. A hilbert scheme in computer vision. *Canadian Journal of Mathematics*, 65, 07 2011.
5. Yulia Alexandr and Alexander Heaton. Logarithmic Voronoi cells. *Algebr. Stat.*, 12(1):75–95, 2021.
6. Carlos Améndola, Lukas Gustafsson, Kathlén Kohn, Orlando Marigliano, and Anna Seigal. The maximum likelihood degree of linear spaces of symmetric matrices. *Matematiche (Catania)*, 76(2):535–557, 2021.
7. Raman Arora, Amitabh Basu, Poorya Mianjy, and Anirbit Mukherjee. Understanding deep neural networks with rectified linear units. In *International Conference on Learning Representations*, 2018.
8. Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *International Conference on Machine Learning*, pages 244–253. PMLR, 2018.
9. S. Asmussen and P. W. Glynn. *Stochastic Simulation: Algorithms and Analysis*. Springer, 2007.
10. Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.
11. Jiakang Bao, Yang-Hui He, Edward Hirst, Johannes Hofscheier, Alexander Kasprzyk, and Suvajit Majumder. Hilbert series, machine learning, and applications to physics. *Physics Letters B*, 827:136966, 2022.
12. Alfred Barnard Basset. *An elementary treatise on cubic and quartic curves*. Cambridge, Deighton, Bell, 1901.
13. S. Basu and A. Lerario. Hausdorff approximations and volume of tubes of singular algebraic sets. *arXiv:2104.05053*, 2021.
14. D. J. Bates, P. Breiding, T. Chen, J. D. Hauenstein, A. Leykin, and F. Ottobre. Numerical nonlinear algebra. *arXiv:2302.08585*, 2023.
15. Daniel Bates, Jonathan Hauenstein, Andrew Sommese, and Charles Wampler. *Numerically solving polynomial systems with Bertini*, volume 25 of *Software, Environments, and Tools*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2013.
16. Adrian Becedas, Kathlén Kohn, and Lorenzo Venturello. Voronoi diagrams of algebraic varieties under polyhedral norms. *arXiv:2209.11463*.
17. Valentina Beorchia, Francesco Galuppi, and Lorenzo Venturello. Eigenschemes of ternary tensors. *SIAM Journal on Applied Algebra and Geometry*, 5(4):620–650, 2021.
18. Edgar A Bernal, Jonathan D Hauenstein, Dhagash Mehta, Margaret H Regan, and Tingting Tang. Machine learning the real discriminant locus. *Journal of Symbolic Computation*, 115:409–426, 2023.
19. D. N. Bernstein. The number of roots of a system of equations. *Funkcional. Anal. i Prilozhen.*, 9(3):1–4, 1975.
20. D. N. Bernstein, A. G. Kušnirenko, and A. G. Hovanskii. Newton polyhedra. *Uspehi Mat. Nauk*, 31(3(189)):201–202, 1976.
21. J. Bezanson, A. Edelman, S. Karpinski, and V. Shah. Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98, 2017.
22. L. Blum, F. Cucker, M. Shub, and S. Smale. *Complexity and real computation*. Springer-Verlag, New York, 1998.
23. Tobias Boege, Jane Coons, Chris Eur, Aida Maraj, and Frank Röttger. Reciprocal maximum likelihood degrees of Brownian motion tree models. *Matematiche (Catania)*, 76(2):383–398, 2021.
24. Viktoriia Borovik and Paul Breiding. A short proof for the parameter continuation theorem. *arXiv:2302.14697*, 2023.
25. Madeline Brandt and Madeleine Weinstein. Voronoi cells in metric algebraic geometry of plane curves. *arXiv:1906.11337*.

26. P. Breiding and O. Marigliano. Random points on an algebraic manifold. *SIAM Journal on Mathematics of Data Science*, 2(3):683–704, 2020.
27. P. Breiding, K. Ranestad, and M. Weinstein. Enumerative geometry of curvature of algebraic hypersurfaces. *arXiv:2206.09130*, 2022.
28. P. Breiding, K. Rose, and S. Timme. Certifying zeros of polynomial systems using interval arithmetic. *ACM Trans. Math. Softw.*, 49(1), mar 2023.
29. P. Breiding and S. Timme. HomotopyContinuation.jl: A Package for Homotopy Continuation in Julia. In *Mathematical Software – ICMS 2018*, pages 458–465, Cham, 2018. Springer International Publishing.
30. P. Breiding and N. Vannieuwenhoven. The condition number of Riemannian approximation problems. *SIAM Journal on Optimization*, 31(1):1049–1077, 2021.
31. Paul Breiding, Sara Kališnik, Bernd Sturmfels, and Madeleine Weinstein. Learning algebraic varieties from samples. *Rev. Mat. Complut.*, 31(3):545–593, 2018.
32. Paul Breiding, Julia Lindberg, Wern Juin Gabriel Ong, and Linus Sommer. Real circles tangent to 3 conics. *Le Matematiche (to appear)*, 2023.
33. Paul Breiding, Kemal Rose, and Sascha Timme. Certifying zeros of polynomial systems using interval arithmetic. *ACM Trans. Math. Software*, 49(1):Art. 11, 14, 2023.
34. Paul Breiding and Sascha Timme. Homotopycontinuation.jl: A package for homotopy continuation in julia. *Math. Software – ICMS 2018*, 458–465, Springer, 2018.
35. Joan Bruna, Kathlén Kohn, and Matthew Trager. Pure and spurious critical points: a geometric study of linear networks. *Internat. Conf. on Learning Representations*, 2020.
36. P. Bürgisser and F. Cucker. *Condition: The Geometry of Numerical Algorithms*, volume 349 of *Grundlehren der mathematischen Wissenschaften*. Springer, Heidelberg, 2013.
37. P. Bürgisser, F. Cucker, and M. Lotz. The probability that a slightly perturbed numerical analysis problem is difficult. *Math. Comput.*, 77:1559–1583, 09 2008.
38. Peter Bürgisser. Condition of intersecting a projective variety with a varying linear subspace. *SIAM Journal on Applied Algebra and Geometry*, 1(1):111–125, 2017.
39. Freddy Cachazo, Nick Early, Alfredo Guevara, and Sebastian Mizera. Scattering equations: from projective spaces to tropical Grassmannians. *J. High Energy Phys.*, (6):039, 32, 2019.
40. Freddy Cachazo, Bruno Umbert, and Yong Zhang. Singular solutions in soft limits. *J. High Energy Phys.*, (5):148, 32, 2020.
41. D. Cartwright and B. Sturmfels. The number of eigenvalues of a tensor. *Linear Algebra and its Applications*, 438(2):942–952, 2013. Tensors and Multilinear Algebra.
42. Jean-Dominique Cassini. *De l'Origine et du progrès de l'astronomie et de son usage dans la géographie et dans la navigation*. L'Imprimerie Royale, 1693.
43. Fabrizio Catanese, Serkan Hoşten, Amit Khetan, and Bernd Sturmfels. The maximum likelihood degree. *Amer. J. Math.*, 128(3):671–697, 2006.
44. Eduardo Cattani, Maria Cueto, Alicia Dickenstein, Sandra Di Rocco, and Bernd Sturmfels. Mixed discriminants. *Mathematische Zeitschrift*, 274, 12 2011.
45. Türkü Özlüm Çelik, Asgar Jammeshan, Guido Montúfar, Bernd Sturmfels, and Lorenzo Venturello. Optimal transport to a variety. In *Mathematical aspects of computer and information sciences*, volume 11989 of *Lecture Notes in Comput. Sci.*, pages 364–381. Springer, Cham, 2020.
46. Türkü Özlüm Çelik, Asgar Jammeshan, Guido Montúfar, Bernd Sturmfels, and Lorenzo Venturello. Wasserstein distance to independence models. *J. Symbolic Comput.*, 104:855–873, 2021.
47. Michel Chasles. *Aperçu historique sur l'origine et le développement des méthodes en géométrie: particulièrement de celles qui se rapportent à la géométrie moderne, suivi d'un mémoire de géométrie sur deux principes généraux de la science, la dualité et l'homographie*. M. Hayez, 1837.
48. F. Chazal and A. Lieutier. The “ $\lambda$ -medial axis”. *Graphical Models*, 67(4):304–331, 2005.
49. Heng-Yu Chen, Yang-Hui He, Shailesh Lal, and Suvajit Majumder. Machine learning lie structures & applications to physics. *Physics Letters B*, 817:136297, 2021.
50. L. Chiantini, G. Ottaviani, and N. Vannieuwenhoven. An algorithm for generic and low-rank specific identifiability of complex tensors. *SIAM J. Matrix Anal. Appl.*, 35(4):1265–1287, 2014.
51. Diego Cifuentes, Corey Harris, and Bernd Sturmfels. The geometry of SDP-exactness in quadratic optimization. *Math. Program.*, 182(1-2, Ser. A):399–428, 2020.
52. Diego Cifuentes, Kristian Ranestad, Bernd Sturmfels, and Madeleine Weinstein. Voronoi cells of varieties. *J. Symbolic Comput.*, 109:351–366, 2022.
53. Roger Cotes. *Harmonia mensurarum*. Robert Smith, 1722.
54. David Cox, John Little, and Donal O’Shea. *Ideals, Varieties, and Algorithms: An introduction to computational algebraic geometry and commutative algebra*. Undergraduate Texts in Math. Springer, Cham, fourth edition, 2015.
55. V. de Silva and L.-H. Lim. Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1084–1127, 2008.

56. Michel Demazure. Sur deux problemes de reconstruction. Technical Report 882, INRIA, 1988.
57. J. W. Demmel. On condition numbers and the distance to the nearest ill-posed problem. *Numer. Math.*, 51(3):251–289, 1987.
58. S. Di Rocco, D. Eklund, and M. Weinstein. The bottleneck degree of algebraic varieties. *SIAM Journal on Applied Algebra and Geometry*, 4(1):227–253, 2020.
59. Sandra Di Rocco, David Eklund, and Madeleine Weinstein. The bottleneck degree of algebraic varieties. *SIAM J. Appl. Algebra Geom.*, 4(1):227–253, 2020.
60. M. do Carmo. *Riemannian Geometry*. Birkhäuser, 1993.
61. Michael Douglas, Subramanian Lakshminarayanan, and Yidi Qi. Numerical calabi-yau metrics from holomorphic networks. In *Mathematical and Scientific Machine Learning*, pages 223–252. PMLR, 2022.
62. Jan Draisma, Emil Horobet, Giorgio Ottaviani, Bernd Sturmfels, and Rekha R. Thomas. The Euclidean distance degree of an algebraic variety. *Found. Comput. Math.*, 16(1):99–149, 2016.
63. Jan Draisma and Jose Rodriguez. Maximum likelihood duality for determinantal varieties. *Int. Math. Res. Not. IMRN*, (20):5648–5666, 2014.
64. Simon S Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. *Advances in neural information processing systems*, 31, 2018.
65. Eliana Duarte, Orlando Marigliano, and Bernd Sturmfels. Discrete statistical models with rational maximum likelihood estimator. *Bernoulli*, 27(1):135–154, 2021.
66. Timothy Duff, Kathlén Kohn, Anton Leykin, and Tomas Pajdla. PLMP – Point-Line Minimal Problems in Complete Multi-View Visibility. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1675–1684, 2019.
67. E. Dufresne, P. Edwards, H. Harrington, and J. Hauenstein. Sampling real algebraic varieties for topological data analysis. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 1531–1536, 2019.
68. Emily Dufresne, Parker Edwards, Heather Harrington, and Jonathan Hauenstein. Sampling real algebraic varieties for topological data analysis. *18th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2019.
69. A. Edelman and E. Kostlan. How many zeros of a random polynomial are real? *Math. Soc. Mathematical Reviews*, 32:1–37, 05 1995.
70. David Eklund. The numerical algebraic geometry of bottlenecks. *Adv. in Appl. Math.*, 142:Paper No. 102416, 20, 2023.
71. Chris Eur, Tara Fife, Jose Samper, and Tim Seynnaeve. Reciprocal maximum likelihood degrees of diagonal linear concentration models. *Matematiche (Catania)*, 76(2):447–459, 2021.
72. S. Friedland and G. Ottaviani. The number of singular vector tuples and uniqueness of best rank-one approximation of tensors. *Foundations of Computational Mathematics*, 14(6):1209–1242, 2014.
73. William Fulton. *Intersection theory*, volume 2 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1984.
74. William Fulton. *Intersection theory*. Springer-Verlag, Berlin, 1998.
75. William Fulton, Steven Kleiman, and Robert MacPherson. About the enumeration of contacts. *Algebraic Geometry – Open Problems*, 997:156–196, 1983.
76. Israel Gel’fand, Mikhail Kapranov, and Andrei Zelevinsky. *Discriminants, resultants, and multidimensional determinants*. Mathematics: Theory & Applications. Birkhäuser Boston, Inc., Boston, MA, 1994.
77. Israel M Gelfand, Mikhail M Kapranov, and Andrei V Zelevinsky. *Discriminants, resultants, and multidimensional determinants*. Birkhäuser, 1994.
78. A. Gray. *Tubes*. Birkhäuser Verlag, Basel, second edition, 2004.
79. Dan Grayson and Michael Stillman. Macaulay2, a software system for research in algebraic geometry. available at [www.math.uiuc.edu/Macaulay2/](http://www.math.uiuc.edu/Macaulay2/).
80. J Elisenda Grigsby, Kathryn Lindsey, Robert Meyerhoff, and Chenxi Wu. Functional dimension of feedforward relu neural networks. *arXiv:2209.04036*, 2022.
81. Philipp Grohs and Gitta Kutyniok. *Mathematical Aspects of Deep Learning*. Cambridge University Press, 2022.
82. Richard Hartley. Projective reconstruction and invariants from multiple images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(10):1036–1041, 1994.
83. Richard Hartley and Frederik Schaffalitzky. Reconstruction from projections using Grassmann tensors. *International Journal of Computer Vision*, 83(3):274–293, 2009.
84. Richard Hartley and Peter Sturm. Triangulation. *Computer vision and image understanding*, 68(2):146–157, 1997.
85. Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge, 2nd edition, 2003.
86. J. D. Hauenstein and D. Kahle. Stochastic exploration of real varieties via variety distributions. Preprint available at [www.nd.edu/~jhauenst/preprints/khVarietyDistributions.pdf](http://www.nd.edu/~jhauenst/preprints/khVarietyDistributions.pdf), 2023.
87. J. D. Hauenstein and F. Sottile. Algorithm 921: alphaCertified: certifying solutions to polynomial systems. *ACM Trans. Math. Software*, 38(4):Art. 28, 20, 2012.

88. Jonathan Hauenstein, Jose Israel Rodriguez, and Bernd Sturmfels. Maximum likelihood for matrices with rank constraints. *J. Algebr. Stat.*, 5(1):18–38, 2014.
89. Kathryn Heal, Avinash Kulkarni, and Emre Can Sertöz. Deep learning Gauss–Manin connections. *Advances in Applied Clifford Algebras*, 32(2):24, 2022.
90. Didier Henrion, Jean Bernard Lasserre, and Carlo Savorgnan. Approximate volume and integration for basic semialgebraic sets. *SIAM Rev.*, 51(4):722–743, 2009.
91. Otto Hesse. Die cubische Gleichung, von welcher die Lösung des Problems der Homographie von M. Chasles abhängt. *Journal für die reine und angewandte Mathematik*, 62, 1863.
92. Anders Heyden and Kalle Åström. Algebraic properties of multilinear constraints. *Mathematical Methods in the Applied Sciences*, 20(13):1135–1162, 1997.
93. N. J. Higham. *Accuracy and stability of numerical algorithms*. Society for Industrial and Applied Mathematics, second edition, 1996.
94. Serkan Hoşten, Amit Khetan, and Bernd Sturmfels. Solving the likelihood equations. *Found. Comput. Math.*, 5(4):389–407, 2005.
95. Audun Holme. The geometric and numerical properties of duality in projective algebraic geometry. *Manuscripta mathematica*, 61:145–162, 1988.
96. Emil Horobet and Madeleine Weinstein. Offset hypersurfaces and persistent homology of algebraic varieties. *Comput. Aided Geom. Design*, 74:101767, 14, 2019.
97. E. Horobet and M. Weinstein. Offset hypersurfaces and persistent homology of algebraic varieties. *Computer Aided Geometric Design*, 74:101767, 2019.
98. D. Hough. *Explaining and Ameliorating the Condition of Zeros of Polynomials*. PhD Thesis, Mathematics Department, University of California, Berkeley, 1977.
99. R. Howard. The kinematic formula in Riemannian homogeneous spaces. *Mem. Amer. Math. Soc.*, 106(509):vi+69, 1993.
100. Petr Hrubař, Timothy Duff, Anton Leykin, and Tomas Pajdla. Learning to solve hard minimal problems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5532–5542, 2022.
101. Kun Huang, Robert Fossum, and Yi Ma. Generalized rank conditions in multiple view geometry with applications to dynamical scenes. In *European Conference on Computer Vision*, pages 201–216. Springer, 2002.
102. Zongyan Huang, Matthew England, David J Wilson, James Bridge, James H Davenport, and Lawrence C Paulson. Using machine learning to improve cylindrical algebraic decomposition. *Mathematics in Computer Science*, 13:461–488, 2019.
103. Birkett Huber and Bernd Sturmfels. A polyhedral method for solving sparse polynomial systems. *Math. Comp.*, 64(212):1541–1555, 1995.
104. June Huh. The maximum likelihood degree of a very affine variety. *Compos. Math.*, 149(8):1245–1266, 2013.
105. June Huh. Varieties with maximum likelihood degree one. *J. Algebr. Stat.*, 5(1):1–17, 2014.
106. June Huh and Bernd Sturmfels. Likelihood geometry. In *Combinatorial algebraic geometry*, volume 2108 of *Lecture Notes in Math.*, pages 63–117. Springer, Cham, 2014.
107. Kenji Kawaguchi. Deep learning without poor local minima. *Advances in neural information processing systems*, 29, 2016.
108. Achim Kehrein and Martin Kreuzer. Characterizations of border bases. *Journal of Pure and Applied Algebra*, 196(2):251–270, 2005.
109. Joe Kileel and Kathlén Kohn. Snapshot of algebraic vision. *arXiv:2210.11443*, 2022.
110. Joe Kileel, Matthew Trager, and Joan Bruna. On the expressive power of deep polynomial neural networks. *Advances in neural information processing systems*, 32, 2019.
111. Steven Kleiman. Tangency and duality. In *Proceedings 1984 Vancouver Conference in Algebraic Geometry*, pages 163–226. Amer. Math. Soc., 1986.
112. F. Klein. Eine neue Relation zwischen den Singularitäten einer algebraischen Curve. *Math. Ann.*, 10(2):199–209, 1876.
113. Kathlén Kohn, Thomas Merkh, Guido Montúfar, and Matthew Trager. Geometry of linear convolutional networks. *SIAM Journal on Applied Algebra and Geometry*, 6(3):368–406, 2022.
114. Kathlén Kohn, Guido Montúfar, Vahid Shahverdi, and Matthew Trager. Function space and critical points of linear convolutional networks. *arXiv:2304.05752*, 2023.
115. Eric Kostlan. On the distribution of roots of random polynomials. In *From Topology to Computation: Proceedings of the Smalefest (Berkeley, CA, 1990)*, pages 419–431. Springer, New York, 1993.
116. Eric Kostlan. On the expected number of real roots of a system of random polynomial equations. In *Foundations of computational mathematics (Hong Kong, 2000)*, pages 149–188. World Sci. Publ., River Edge, NJ, 2002.
117. Christoph Koutschan. Holonomicfunctions: A mathematica package for dealing with multivariate holonomic functions, including closure properties, summation, and integration. available at [www3.risc.jku.at/research/combinat/software/ergosum/RISC/HolonomicFunctions.html](http://www3.risc.jku.at/research/combinat/software/ergosum/RISC/HolonomicFunctions.html).

118. Pierre Lairez. Computing periods of rational integrals. *Math. Comp.*, 85(300):1719–1752, 2016.
119. Pierre Lairez, Marc Mezzarobba, and Mohab Safey El Din. Computing the volume of compact semi-algebraic sets. In *ISSAC’19—Proceedings of the 2019 ACM International Symposium on Symbolic and Algebraic Computation*, pages 259–266. ACM, New York, 2019.
120. Jean Bernard Lasserre. *Moments, positive polynomials and their applications*, volume 1 of *Imperial College Press Optimization Series*. Imperial College Press, London, 2010.
121. L. D. Lathauwer. Decompositions of a higher-order tensor in block terms - Part II: Definitions and uniqueness. *SIAM J. Matrix Anal. Appl.*, 30:1033–1066, 2008.
122. Thomas Laurent and James Brecht. Deep linear networks with arbitrary loss: All local minima are global. In *International Conference on Machine Learning*, pages 2902–2907. PMLR, 2018.
123. J. M. Lee. *Riemannian Manifolds: Introduction to Curvature*. Springer, 1997.
124. J. M. Lee. *Introduction to Smooth Manifolds*. Springer, New York, USA, 2 edition, 2013.
125. K. Lee. Certifying approximate solutions to polynomial systems on Macaulay2. *ACM Communications in Computer Algebra*, 53(2):45–48, 2019.
126. T. Lelievre, G. Stoltz, and W. Zhang. Multiple projection mcmc algorithms on submanifolds. *IMA Journal of Numerical Analysis*, 2022.
127. Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.
128. M. Lotz. On the volume of tubular neighborhoods of real algebraic varieties. *Proceedings of the American Mathematical Society*, 143, 10 2012.
129. Laurent Manivel, Mateusz Michałek, Leonid Monin, Tim Seynnaeve, and Martin Vodička. Complete quadrics: Schubert calculus for gaussian models and semidefinite programming. *Journal of the European Mathematical Society*, 2023.
130. Laurentiu G Maxim, Jose I Rodriguez, and Botong Wang. Euclidean distance degree of the multiview variety. *SIAM Journal on Applied Algebra and Geometry*, 4(1):28–48, 2020.
131. Clerk Maxwell. On the description of oval curves, and those having a plurality of foci. *Proceedings of the Royal Society of Edinburgh*, 2, 1846.
132. Dhagash Mehta, Tianran Chen, Tingting Tang, and Jonathan D Hauenstein. The loss surface of deep linear networks viewed through the algebraic geometry lens. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5664–5680, 2021.
133. S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Springer-Verlag, London., 1993.
134. Mateusz Michałek, Leonid Monin, and Jarosław Wiśniewski. Maximum likelihood degree, complete quadrics, and  $\mathbb{C}^*$ -action. *SIAM J. Appl. Algebra Geom.*, 5(1):60–85, 2021.
135. Mateusz Michałek and Bernd Sturmfels. *Invitation to nonlinear algebra*, volume 211 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2021.
136. Guido Montúfar, Yue Ren, and Leon Zhang. Sharp bounds for the number of regions of maxout networks and vertices of minkowski sums. *SIAM Journal on Applied Algebra and Geometry*, 6(4):618–649, 2022.
137. A. P. Morgan and A. J. Sommese. Coefficient-parameter polynomial continuation. *Applied Mathematics and Computation*, 29(2):123–160, 1989.
138. Gabin Maxime Nguegnang, Holger Rauhut, and Ulrich Terstiege. Convergence of gradient descent for learning linear neural networks. *arXiv:2108.02040*, 2021.
139. Jiawang Nie, Pablo A Parrilo, and Bernd Sturmfels. Semidefinite representation of the k-ellipse. In *Algorithms in Algebraic Geometry*, volume 146 of *I.M.A. Volumes in Mathematics and its Applications*, pages 117–132. Springer, 2008.
140. Luke Oeding. The quadrifocal variety. *Linear Algebra and its Applications*, 512:306–330, 2017.
141. B. O’Neill. *Elementary Differential Geometry*. Elsevier, revised second edition edition, 2001.
142. G. Ottaviani and L. Sodomaco. The distance function from a real algebraic variety. *Computer Aided Geometric Design*, 82:101927, 2020.
143. Giorgio Ottaviani and Luca Sodomaco. The distance function from a real algebraic variety. *Comput. Aided Geom. Design*, 82:101927, 20, 2020.
144. Giorgio Ottaviani, Pierre-Jean Spaenlehauer, and Bernd Sturmfels. Exact solutions in structured low-rank approximation. *SIAM J. Matrix Anal. Appl.*, 35(4):1521–1542, 2014.
145. S. Smale P. Niyogi and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geometry*, (39):419–441, 2008.
146. Dylan Peifer, Michael Stillman, and Daniel Halpern-Leistner. Learning selection strategies in buchberger’s algorithm. In *International Conference on Machine Learning*, pages 7575–7585. PMLR, 2020.
147. P. Petersen. *Riemannian Geometry*. Springer, second edition, 2006.
148. R. Piene, C. Riener, and B. Shapiro. Return of the plane evolute. *arXiv:2110.11691*, 2021.
149. Ragni Piene. Polar classes of singular varieties. In *Annales scientifiques de l’École Normale Supérieure*, volume 11, pages 247–276, 1978.

150. Ragni Piene. Cycles polaires et classes de chern pour les variétés projectives singulières. *Introduction à la théorie des singularités, II*, 37:7–34, 1988.
151. Julius Plücker. *Gesammelte wissenschaftliche Abhandlungen: Im Auftrag des Kgl*, volume 2. BG Teubner, 1895.
152. S. Pokutta and E. Wirth. Conditional gradients for the approximately vanishing ideal. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain.*, 2002.
153. Y. Qi, P. Comon, and L.-H. Lim. Semialgebraic geometry of nonnegative tensor rank. *SIAM J. Matrix Anal. Appl.*, 37:1556–1580, 11 2016.
154. Adityanarayanan Radhakrishnan, Mikhail Belkin, and Caroline Uhler. Overparameterized neural networks implement associative memory. *Proceedings of the National Academy of Sciences*, 117(44):27162–27170, 2020.
155. Christophe Raffalli. Distance to the discriminant. *arXiv:1404.7253*, 2014.
156. J. R. Rice. A theory of condition. *SIAM J. Numer. Anal.*, 3:287–310, 1966.
157. Gareth O. Roberts and Jeffrey S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71, 2004.
158. S. Di Rocco, D. Eklund, and O. Gäfvert. Sampling and homology via bottlenecks. *Math. Comp.*, 91(338):2969–2995, 2022.
159. Jose Israel Rodriguez and Botong Wang. The maximum likelihood degree of mixtures of independence models. *SIAM J. Appl. Algebra Geom.*, 1(1):484–506, 2017.
160. Mutsumi Saito, Bernd Sturmfels, and Nobuki Takayama. *Gröbner deformations of hypergeometric differential equations*, volume 6 of *Algorithms and Computation in Mathematics*. Springer-Verlag, Berlin, 2000.
161. George Salmon. *A Treatise on the Higher Plane Curves*. Hodges and Smith, Dublin, 1852.
162. George Salmon. *A Treatise on the Analytic Geometry of Three Dimensions*. Hodges, Smith, and Company, 1865.
163. Muhamad Risqi U Saputra, Andrew Markham, and Niki Trigoni. Visual slam and structure from motion in dynamic environments: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–36, 2018.
164. Anna-Laura Sattelberger and Bernd Sturmfels. D-modules and holonomic functions. *to appear in: Varieties, Polyhedra and Computation, EMS Series of Congress Reports*, arxiv:1910.01395.
165. Luca Sodomaco. The distance function from the variety of partially symmetric rank-one tensors. *PhD thesis, Università degli Studi di Firenze*, 2020.
166. A. J. Sommese and C. W. Wampler. *The Numerical Solution of Systems of Polynomials Arising in Engineering and Science*. World Scientific, 2005.
167. Henrik Stewénius, Frederik Schaffalitzky, and David Nistér. How hard is 3-view triangulation really? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 686–693, 2005.
168. Elisabetta Strickland. On the conormal bundle of the determinantal variety. *J. Algebra*, 75(2):523–537, 1982.
169. Rud Sturm. Das Problem der Projectivität und seine Anwendung auf die Flächen zweiten Grades. *Mathematische Annalen*, 1(4):533–574, 1869.
170. Bernd Sturmfels. *Solving Systems of Polynomial Equations*. Number 97 in CBMS Regional Conferences Series. American Mathematical Society, 2002.
171. Bernd Sturmfels and Simon Telen. Likelihood equations and scattering amplitudes. *Algebr. Stat.*, 12(2):167–186, 2021.
172. Bernd Sturmfels, Sascha Timme, and Piotr Zwiernik. Estimating linear covariance models with numerical nonlinear algebra. *Algebr. Stat.*, 11(1):31–52, 2020.
173. Seth Sullivant. *Algebraic statistics*, volume 194 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2018.
174. Gyula Sz.-Nagy. Tschirnhaus' sche eiflächen und eikurven. *Acta Mathematica Hungarica*, pages 36–45, 1950.
175. Matteo Tacchi, Jean Bernard Lasserre, and Didier Henrion. Stokes, Gibbs, and volume computation of semi-algebraic sets. *Discrete Comput. Geom.*, 69(1):260–283, 2023.
176. Matteo Tacchi, Tillmann Weisser, Jean Bernard Lasserre, and Didier Henrion. Exploiting sparsity for semi-algebraic set volume computation. *Found. Comput. Math.*, 22(1):161–209, 2022.
177. Gerald J Toomer. *Apollonius: Conics Books V to VII: The Arabic Translation of the Lost Greek Original in the Version of the Banū Mūsā*, volume 1. Springer, 1990.
178. L. N. Trefethen and D. Bau. *Numerical Linear Algebra*. SIAM, 1997.
179. A. Turing. Rounding-off errors in matrix processes. *The Quarterly Journal of Mechanics and Applied Mathematics*, 1, 1948.
180. H. Weyl. On the volume of tubes. *Amer. J. Math.*, 61(2):461–472, 1939.
181. J. H. Wilkinson. Note on matrices with a very ill-conditioned eigenproblem. *Numer. Math.*, 19:176–178, 1972.
182. Francis Williams, Matthew Trager, Daniele Panozzo, Claudio Silva, Denis Zorin, and Joan Bruna. Gradient dynamics of shallow univariate relu networks. *Advances in Neural Information Processing Systems*, 32, 2019.
183. Lior Wolf and Amnon Shashua. On projection matrices  $\mathbb{P}^k \rightarrow \mathbb{P}^2$ ,  $k = 3, \dots, 6$ , and their applications in computer vision. *International Journal of Computer Vision*, 48(1):53–67, 2002.
184. Doron Zeilberger. A holonomic systems approach to special functions identities. *J. Comput. Appl. Math.*, 32(3):321–368, 1990.

185. Liwen Zhang, Gregory Naitzat, and Lek-Heng Lim. Tropical geometry of deep neural networks. In *International Conference on Machine Learning*, pages 5824–5832. PMLR, 2018.
186. Qunjie Zhou, Torsten Sattler, Marc Pollefeys, and Laura Leal-Taixe. To learn or not to learn: Visual localization from essential matrices. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3319–3326. IEEE, 2020.