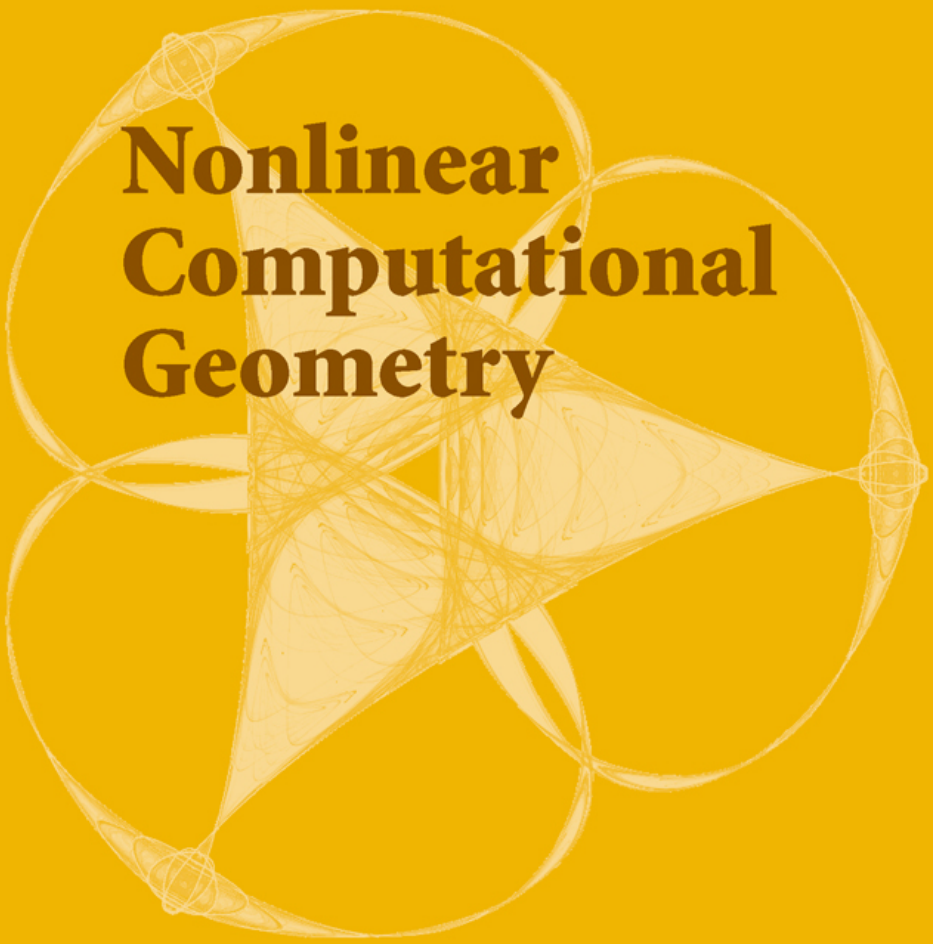
THE IMA VOLUMES IN MATHEMATICS AND ITS APPLICATIONS

EDITORS Ioannis Z. Emiris
Frank Sottile
Thorsten Theobald



Nonlinear Computational Geometry

Springer

The IMA Volumes in Mathematics and its Applications

Volume 151

Series Editors
Fadil Santosa Markus Keel

For other titles published in this series, go to
www.springer.com/series/811

Ioannis Z. Emiris Frank Sottile
Thorsten Theobald
Editors

Nonlinear Computational Geometry

Springer

Editors
Ioannis Z. Emiris
Lab of Geometric & Algebraic Algorithms
Department of Informatics
& Telecommunications
National and Kapodistrian University
of Athens
Panepistimiopolis, 15784
Greece
http://cgi.di.uoa.gr/~emiris/index-eng.html

Thorsten Theobald
FB 12 - Institut für Mathematik
Johann Wolfgang Goethe-Universität
Robert-Mayer-Str. 10
D-60325 Frankfurt am Main
Germany
http://www.math.uni-frankfurt.de/~theobald/

Frank Sottile
Department of Mathematics
Texas A&M University
College Station, TX 77843
USA
http://www.math.tamu.edu/~sottile

Series Editors
Fadil Santosa
Markus Keel
Institute for Mathematics
& its Applications
University of Minnesota
Minneapolis, MN 55455
USA

ISSN 0940-6573
ISBN 978-1-4419-0998-5 e-ISBN 978-1-4419-0999-2
DOI 10.1007/978-1-4419-0999-2
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 200931559

Mathematics Subject Classification (2000): 14Q, 68U05, 68W30, 65D, 52A35

Camera-ready copy provided by the IMA.

Springer is part of Springer Science+Business Media (www.springer.com)

FOREWORD

This IMA Volume in Mathematics and its Applications

NONLINEAR COMPUTATIONAL GEOMETRY

contains papers presented at a highly successful one-week workshop held on May 29–June 2, 2007 on the same title. The event was an integral part of the 2006–2007 IMA Thematic Year on "Applications of Algebraic Geometry." We are grateful to all the participants for making this workshop a very productive and stimulating event.

We owe special thanks to Ioannis Z. Emiris (Department of Informatics and Telecommunications, National and Kapodistrian University of Athens), Frank Sottile (Department of Mathematics, Texas A&M University), and Thorsten Theobald (Institut für Mathematik, Johann Wolfgang Goethe-Universität) for their superb role as workshop organizers and editors of these proceedings. We also thank Ron Goldman (Department of Computer Science, Rice University) for his valuable contribution in organizing the workshop.

We take this opportunity to thank the National Science Foundation for its support of the IMA.

Series Editors

Fadil Santosa, Director of the IMA

Markus Keel, Deputy Director of the IMA

PREFACE

An original goal of *algebraic geometry* was to understand curves and surfaces in three dimensions. From these roots, algebraic geometry has grown into a theoretically deep and technically sophisticated field. Recently, questions from robotics, computer vision, computer-aided geometric design and molecular biology, together with the development of computational methods, have brought these original questions back to the forefront of research. The implicitization of parametric surfaces, the geometry of molecules, mechanical design and computer vision all lead to problems that are challenging from the perspective of computational algebraic geometry.

In recent decades, *computational geometry* has developed as a discipline at the intersection of mathematics and computer science that provides effective and algorithmic methods for treating geometric problems. For natural reasons, the primary focus in computational geometry has been on polyhedral (piecewise-linear) problems.

The challenge has arisen to combine the applicable methods of algebraic geometry with the proven techniques of piecewise-linear computational geometry (such as Voronoi diagrams and hyperplane arrangements) to generalize the latter to curved objects. These research efforts may be summarized under the term *nonlinear computational geometry.* In this area, the development of reliable and practical algorithms is often based on interrelated techniques that incorporate both symbolic and numerical elements.

Within the thematic year *Applications of Algebraic Geometry* 2006/ 2007 at the Institute for Mathematics and its Applications in Minneapolis, a week-long workshop was devoted to this topic of nonlinear computational geometry. This workshop took place from May 29 to June 2, 2007, and was organized by I.Z. Emiris, R. Goldman, F. Sottile, and T. Theobald. Around 100 experts in this emerging field attended.

The present volume is comprised of nine contributions covering the spectrum of topics from the workshop. Its purpose is to establish a collection of research and expository articles describing the state-of-the-art in nonlinear computational geometry and the challenges it poses for computational geometry, algebraic geometry, and geometric modeling.

In the first chapter, *Spectral techniques to explore point clouds in Euclidean space with applications in structural biology, Frédéric Cazals, Frédéric Chazal and Joachim Giesen* survey recent progress in spectral techniques for machine learning, such as the principal component analysis (PCA) and multi-dimensional scaling (MDS), in order to identify structure in point clouds. The authors then offer an overview of such methods applied to understanding the geometry of large molecules and, in particular, the important open problem of protein folding in bioinformatics.

Algebraic hypersurfaces in space may have dual descriptions in terms of an implicit equation or via a parametric representation. Transforming one of these representations into the other and understanding their connection is an ubiquitous task in nonlinear geometric computations. In the chapter *Rational parameterizations, intersection theory and Newton polytopes*, *Carlos D'Andrea and Martín Sombra* exploit recent advances in algebraic geometry and combinatorial geometry to describe the Newton polytope and the support of the implicit equation of a given rational parametric hypersurface, by means of invariants of its parameterization.

When transferring classical problems from the space $\mathbb{R}^n$ to the space of lines, i.e., to the Grassmannian manifold, problems and structures obtain an additional nonlinear component. In the chapter *Some discrete properties of the space of line transversals to disjoint balls*, *Xavier Goaoc* describes the state-of-the-art in generalizing classical statements from convex geometry (in particular Helly's Theorem and geometric transversals) to the space of lines, and shows how these can lead to algorithmic applications.

Many problems in kinematics have natural formulations in terms of polynomial equations, and already more than a century ago these formulations were studied and applied to kinematics. In *Algebraic geometry and kinematics*, *Manfred Husty* and *Peter Schröcker* explain how the classical Study mapping can be used to transform kinematic problems into algebraic-geometric problems and discuss the analysis of mechanisms from a modern point of view.

In geometric modeling of real-world phenomena, offsets—surfaces at a constant normal distance to a given surface—frequently arise. Even when the original surface is rational, its offsets often are not, and it is a challenge to understand those surfaces with rational offsets. *Rimvydas Krasauskas* and *Martin Peternell* survey this in *Rational offset surfaces and their modeling applications*. A particular focus is given on Pythagorean normal surfaces as well as on the viewpoint afforded by Laguerre geometry.

In his contribution *A list of challenges for real algebraic plane curve visualization software*, *Oliver Labs* discusses the issue of correctly visualizing a real plane algebraic curve. The occurrence of singularities can make this task quite challenging. Particularly difficult are those curves which contain "complicated" singularities such as high tangencies or many halfbranches. The exposition describes several classes of curves to serve as benchmarks for future visualization software.

In the chapter *Subdivision method for arrangement computation of semi-algebraic curves*, *Bernard Mourrain and Julien Wintz* present a synthesis of existing approaches, enhanced with new tools, to compute arrangements of semi-algebraic sets in a certified and efficient manner. Their approach is to apply a subdivision technique and to analyze the topological structure through this process. These concepts and methods are illustrated by an implementation in the geometric modeler Axel.

Solving problems in nonlinear computational geometry exactly often

leads to geometric predicates of high algebraic degree. In the chapter *Invariant-based characterization of the relative position of two projective conics*, *Sylvain Petitjean* studies these predicates for the fundamental problem of characterizing the relative position of two given conics. For his bounds, he applies methods of classical invariant theory.

In the architectural design of glass/steel panel structures, free-form surfaces may be approximated by polyhedral surfaces with hexagonal facets. The final chapter, *A note on planar hexagonal meshes*, *Wenping Wang and Yang Liu* first describes the applications and the theory of such surfaces. It then discusses different algorithms for generating hexagonal meshes, including a new one proposed by the authors.

We hope that the reader will enjoy the articles in the volume and that the articles offer the reader an overview of the current developments in nonlinear computational geometry.

Acknowledgments: The editors wish to thank the Institute for Mathematics and its Applications for providing an inspiring and fruitful workshop atmosphere as well as for their help in editing the volume.

Ioannis Z. Emiris
Department of Informatics & Telecommunications
National and Kapodistrian University of Athens
http://cgi.di.uoa.gr/~emiris/index-eng.html

Frank Sottile
Department of Mathematics
Texas A&M University
http://www.math.tamu.edu/~sottile

Thorsten Theobald
Institut für Mathematik
Johann Wolfgang Goethe-Universität
http://www.math.uni-frankfurt.de/~theobald/

CONTENTS

Foreword v

Preface vii

Spectral techniques to explore point clouds in Euclidean space, with applications to collective coordinates in structural biology 1
Frédéric Cazals, Frédéric Chazal, and Joachim Giesen

Rational parametrizations, intersection theory, and Newton polytopes 35
Carlos D'Andrea and Martín Sombra

Some discrete properties of the space of line transversals to disjoint balls 51
Xavier Goaoc

Algebraic geometry and kinematics 85
Manfred L. Husty and Hans-Peter Schröcker

Rational offset surfaces and their modeling applications 109
Rimvydas Krasauskas and Martin Peternell

A list of challenges for real algebraic plane curve visualization software 137
Oliver Labs

A subdivision method for arrangement computation of semi-algebraic curves 165
Bernard Mourrain and Julien Wintz

Invariant-based characterization of the relative position of two projective conics 189
Sylvain Petitjean

A note on planar hexagonal meshes 221
Wenping Wang and Yang Liu

List of workshop participants 235

SPECTRAL TECHNIQUES TO EXPLORE POINT CLOUDS IN EUCLIDEAN SPACE, WITH APPLICATIONS TO COLLECTIVE COORDINATES IN STRUCTURAL BIOLOGY

FRÉDÉRIC CAZALS*, FRÉDÉRIC CHAZAL†, AND JOACHIM GIESEN‡

Abstract. Life sciences, engineering, or telecommunications provide numerous systems whose description requires a large number of variables. Developing insights into such systems, forecasting their evolution, or monitoring them is often based on the inference of correlations between these variables. Given a collection of points describing states of the system, questions such as inferring the effective number of independent parameters of the system (its intrinsic dimensionality) and the way these are coupled are paramount to develop models. In this context, this paper makes two contributions.

First, we review recent work on spectral techniques to organize point clouds in Euclidean space, with emphasis on the main difficulties faced. Second, after a careful presentation of the bio-physical context, we present applications of dimensionality reduction techniques to a core problem in structural biology, namely protein folding.

Both from the computer science and the structural biology perspective, we expect this survey to shed new light on the importance of *non linear computational geometry* in geometric data analysis in general, and for protein folding in particular.

Contents

1 Introduction . . . 2
1.1 Geometric data analysis and spectral point cloud processing . . . 2
1.2 Spectral methods and alternatives . . . 3
1.3 An application in structural biology: Protein folding . . . 5
1.4 Notations and paper overview . . . 6
2 PCA and MDS . . . 6
2.1 PCA . . . 7
2.2 MDS . . . 8
3 Localization . . . 8
3.1 Neighborhood criteria . . . 8
3.2 Dimension detection using PCA . . . 9
4 Turning non-linear . . . 10
4.1 Maximum variance unfolding (MVU) . . . 10
4.2 Locally linear embedding (LLE) . . . 11
4.3 ISOMAP . . . 12
4.4 Laplacian eigenmaps . . . 12
4.5 Hessian eigenmaps (HLLE) . . . 13
4.6 Diffusion maps . . . 14
5 Applications in structural biology: the folding problem . . . 15
5.1 Folding: from experiments to modeling . . . 16
5.2 Energy landscapes and dimensionality reduction . . . 16
5.2.1 Potential and free energy landscape . . . 16
5.2.2 Enthalpy-entropy compensation, energy funnel, ruggedness and frustration . . . 16
5.2.3 Cooperativity and correlated motions . . . 18
5.3 Bio-physics: Pre-requisites . . . 19
5.3.1 Molecular dynamics simulations . . . 19
5.3.2 Models, potential energy landscapes and their ruggedness . . . 19
5.3.3 Morse theory and singularity theory . . . 20
5.3.4 Free energy landscapes and reaction coordinates . . . 20
5.3.5 Folding probability p_{fold} . . . 21
5.4 Inferring reaction coordinates . . . 24
5.4.1 Reaction coordinates? . . . 24
5.4.2 Contacts based analysis . . . 24
5.4.3 Dimension reduction based analysis . . . 26
5.4.4 Morse theory related analysis . . . 27
6 Outlook . . . 28

*INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 Sophia Antipolis, France (Frederic.Cazals@inria.fr).

†INRIA Saclay, Parc Orsay Université, 4 rue Jacques Monod, 91893 Orsay Cedex, France (Frederic.Chazal@inria.fr).

‡Institut fuer Informatik, Ernst-Abbe-Platz 2, D-07743 Jena, Germany (jgiesen@mpi-inf.mpg.de).

I.Z. Emiris et al. (eds.), *Nonlinear Computational Geometry*, The IMA Volumes in Mathematics and its Applications 151, DOI 10.1007/978-1-4419-0999-2_1,

1. Introduction.

1.1. Geometric data analysis and spectral point cloud processing. Modeling the climate, understanding the interplay between proteins, metabolites and nucleic acids making up a regulation network within a cell, or unraveling the connexions between spiking neurons are example problems where a large number of variables interplay in a complex non linear way. Developing insights into such systems, forecasting their evolution, or monitoring them is often based on the inference of correlations between these variables. More precisely, learning such correlations from experiments is paramount to model development, as theory and experimental inference are tightly coupled.

Consider a complex system, and assume we are given a number of observations describing different states of the system. In such a setting, we are interested in the question of inferring the effective number of independent parameters of the system (its intrinsic dimensionality) and the way these are coupled. To meet these challenges, a set of new geometric methods, known as manifold learning, have been developed in the machine learning community mainly over the past decade. These methods are based upon the assumption that the observed data –a point cloud in some n dimensional space, lie on or are close to a submanifold M in $\mathbb{R}^d$.

Naturally, given the variety of situations, one cannot expect a single method to meet all needs. Nevertheless, many of the most popular approaches boil down to *spectral methods*. Note that the term *spectral method* is ambiguous and used differently within different communities, e.g., in numerical methods for partial differential equations it often involves the use of the fast Fourier transform. Here we want to use the term in the sense of data analysis similar as van der Maaten et al. did [51]. That is, for us in a spectral method, a symmetric matrix is derived from the point cloud data and the solution to a given optimization problem can be obtained from the eigenvectors of this matrix. We should mention that the term *spectral method* is also used in mesh processing in the geometric modeling community where the symmetric matrix is obtained from the connectivity of the mesh, see [58] for an overview. The geometric optimization problems that lead to a spectral technique are mostly of a *least squares* nature and include the following classical (and archetypical) problems:

(1) Find the k-dimensional subspace that approximates the point cloud best in a least squares sense.
(2) Find the embedding of the point cloud in k-dimensional space that preserves the distances between the points best possible in a least squares sense.

The first problem is called *principal component analysis (PCA)* as it asks for the principal directions (components) of the data. It essentially is a data quantization technique: every data point gets replaces by its projection onto the best approximating k-dimensional subspace. The loss

incurred by the quantization is the variance of the data in the directions orthogonal the best approximating k-dimensional subspace. As long as this variance is small PCA can also be seen as *denoising* the original data. Many machine learning techniques including clustering, classification and semi-supervised learning [32], but also near neighbor indexing and search can benefit from such a denoising.

The second problem is called *multi-dimensional scaling (MDS)*. An important application of MDS is visualization of the point cloud data: the data points get embedded into two- or three-dimensional space, where they can be directly visualized. The main purpose of visualization is to use the human visual system to get insights into the structure of the point cloud data, e.g., the existence of clusters or—for data points labeled with discrete attributes—relations between this attributes. MDS visualization remains to be a popular tool for point cloud data analysis, but of course a lot of information will get lost (and in general cannot be restored by the human visual system) if the *intrinsic dimension* of the data points is larger than three.

Recently the focus in point cloud data analysis shifted: more emphasis is put on detecting non-linear features in the data, although processing the data for visual inspection still is important. What drives this shift in focus is the insight that most features are based on *local* correlations of the data points, but PCA and MDS both have only a global view on the point cloud data. The shift towards local correlations was pioneered by two techniques called *ISOMAP* [48, 21] and *Locally Linear Embedding (LLE)* [45, 46]. It is important to note that focusing on local correlations does not mean that one loses the global picture: for example the global intrinsic dimension of the data can be estimated from local information, whereas it is often (when the data are embedded non-linearly) not possible to derive this information from a purely global analysis. ISOMAP and LLE and their successors (some of which we will also discuss here) can be used both for the traditional purposes data quantization and data visualization. In general they preserve more information of the data (than PCA and MDS) while achieving a similar quantization error or targeting the same embedding dimension for data visualization, respectively.

1.2. Spectral methods and alternatives.

Advantages of spectral methods. Consider a point cloud P sampled from a manifold M embedded in $\mathbb{R}^d$. In this survey, we focus on a set of quite famous methods following a common thread, as they ultimately resort to spectral analysis. They all intend to find the best embedding of the dataset P into an Euclidean space $\mathbb{R}^k$ with respect to some quadratic constraint reflecting different geometric properties of the underlying manifold M. The embedding of the data that minimizes the quadratic constraint can then be interpreted as the best k-dimensional embedding of the data with respect to the geometric property we aim to preserve. In most cases, the

quadratic minimization problem boils down to a general eigenvalue problem ensuring to find a global minimum. Moreover, the embedding can be found by easy-to-implement polynomial time algorithms.

This provides a substantial advantage over iterative or greedy methods based upon Expectation-Maximization like algorithms that do not provide guarantees of global optimality. In particular, for quite large data sets, the methods we consider still provide results when iterative and greedy methods fail due to complexity issues. Another advantage of "spectral methods" is that the quadratic constraint leads to a measurement of the quality of the embedding[1]. At last, "spectral methods" have been widely used and studied in many applications areas (graph theory, mesh processing [58],...) giving rise to a large amount of efficient theoretical and algorithmic tools that can be used for dimensionality reduction.

Approaches not covered. As our focus is on spectral techniques, a number of dimensionality reduction techniques are not covered in this paper. While the reader might consult [51] for a rather exhaustive catalog, the following comments are in order about the missing classes:

- EM-based methods: a large set of manifold learning algorithms developed in the machine learning community adopt a probabilistic point of view, so as to maximize a likelihood (Self Organizing Maps, Generative Topographic Mapping, Principal curves, etc. See [8] for example.). Some of them, like principal curves [31] or generative topographic mapping [9] for example, aim to fit the data set by a parameterized low dimensional (in general 1 or 2) manifold. They usually assume that the topology of the manifold is known and simple (simple curves, planes, discs) and do not allow to deal with data sampled from more complicated shapes.
- Methods related to the Johnson-Lindenstrauss lemma: the Johnson-Lindenstrauss lemma addresses the dimensionality reduction problem of a general point cloud (not necessarily sampled around a low dimensional manifold) in the perspective of preserving the pairwise distances between the points. An extension to points and flats and algebraic surfaces has been proposed in [1].
- Kernel methods: a number of methods, including some of the methods we shall discuss, can be interpreted in the framework of kernel methods. See [29, 53] for example.
- Methods targeting non manifold shapes: more recently, some geometric inference methods have been developed in the case where the shape underlying the data is not assumed to be a smooth man-

[1] For example, in [48], the quality of the embedding obtained is assessed resorting to the residual variance $\sigma_k(k,d)$ defined by:

$$\sigma_k(k,d) = 1 - R^2(\hat{D}_k, D_d) \tag{1}$$

with $R(\hat{D}_k, D_d)$ the correlation coefficient taken over all entries of matrices $\hat{D}_k$ and D_d. The closer to zero this variance, the better the approximation.

ifold. They lead to promising but preliminary results for dimensionality reduction of general shapes [11, 12].

1.3. An application in structural biology: Protein folding. As an application of dimensionality reduction techniques in general, and of spectral methods in particular, we shall give a detailed account of one of the core open problems in structural biology, namely protein folding: how does a protein reach its folded-, i.e., its biologically active state, from the unfolded one? As of October 2007, about 1,000 genomes have been fully sequenced or are about to be so, while the Protein Data Bank contains (a mere) 40,000 structures. The question of understanding folding so as to predict the structure of a protein from its sequence is therefore central[2], to foster the understanding of central mechanisms in the cell, but also to perform protein engineering with applications ranging from drug design to bio-technologies.

Aside these general incentives, a number of technical ones advocate this particular problem.

First, the question of folding is closely related to a specific d-dimensional manifold which associates an energy to a conformation (the energy landscape), on which point clouds are sampled thanks to simulations techniques like the prototypical molecular dynamics method. Thus, the underlying mathematical structure is a manifold and not a (stratified) complex of arbitrary topology.

Second, assuming the folded and unfolded conformations correspond to (significant) local minima of the energy landscape, the problem is tantamount to understand *transitions* on this landscape, i.e. paths joining these minima. The difficulty of the problem is rooted in two facts: the high-dimensionality of the landscape ($d = 3n$ or $d = 6n$ as argued below, with n the number of atoms), and its complex topography which reflects the complex interactions (forces) between atoms. These intrinsic difficulties call for dimensionality reduction techniques, so as to exhibit a small number of new variables (typically one or two), called the reaction coordinates, accounting for the transition. These coordinates should match the effective *large amplitude - slow frequency* degrees of freedom of the system, thus providing a simplified view of the process, and easing its interpretation. Thus in essence, one wishes to quantize information located on a non linear manifold, while retaining the essential features.

Third, as opposed to a large number of multi-dimensional data sets, folding features a stimulating interplay between modeling and experiments. The point clouds studied in folding are indeed closely related to a number of experiments in bio-physics, so that one can precisely assess the quality and the interest of dimensionality reduction procedures. Example such experimental methods are dynamic NMR, protein engineering (ϕ-value analysis),

[2]At least for proteins consisting of a single polypeptidic chain, as the formation of multimers also poses docking questions.

laser initiated folding, etc. Describing these procedures is clearly beyond the scope of this survey, and the reader is referred to [84, 75] for starting pointers.

1.4. Notations and paper overview. Throughout this paper we will be using the following notations:

P	point cloud
n	number of point is P
d	dimension of the Euclidean space form which the points in P are drawn
k	target dimension.

The paper is organized as follows. Section 2 presents the two archetypical spectral methods used to explore point clouds, namely PCA and MDS. The question of localizing neighborhoods is discussed in Section 3, while methods meant to accommodate non linear geometries are discussed in Section 4. The application of dimensionality reduction techniques to protein folding is discussed in Section 5. To conclude, Section 6 discusses a number of research challenges.

2. PCA and MDS. In the following we assume that the points in P are centered at the origin, i.e., $\sum_{i=1}^{n} p_i = 0$. Note that this can always achieved by a simple translation: let $\bar{p} = \frac{1}{n}\sum_{i=1}^{n} p_i$ and $p'_i = p_i - \bar{p}$, then $\sum_{i=1}^{n} p'_i = 0$.

Principal component analysis (PCA) asks for the k-dimensional subspace of $\mathbb{R}^d$ that approximates the point set P best possible in a least squares sense and projects P onto that subspace, whereas multidimensional scaling (MDS) in its basic form aims for the k-dimensional embedding of P that preserves the pairwise inner products of the points in P best possible in a least squares sense. In both cases k can range from 1 to $d-1$.

Though different in their motivation and objective, PCA and MDS are almost identical in a technical sense: both can be formulated in terms of eigenvectors of some positive semi-definite matrix derived from the point set P, which itself can be written as a $(d \times n)$-matrix as follows:

$$P = \begin{pmatrix} p_{11} & \cdots & p_{n1} \\ \vdots & & \vdots \\ p_{1d} & \cdots & p_{nd} \end{pmatrix},$$

where p_{ij} is the j'th component of the point $p_i \in P$. From the matrix P one canonically derives two positive semi-definite matrices,

(1) the *covariance matrix* $C = PP^T$, and
(2) the *Gram matrix* $G = P^T P$.

The covariance matrix is a $(d \times d)$-matrix and can also be written as $C = \sum_{i=1}^{n} p_i p_i^T$, whereas the Gram matrix has dimension $n \times n$ and can also be

written as $G = \left(p_i^T p_j\right)$. Both matrices are intimately linked also via their eigenvectors and eigenvalues. We have the following observation.

OBSERVATION 1. *The matrices C and G have the same non-zero (positive) eigenvalues (and thus the same rank).*

Proof. Let $v \in \mathbb{R}^d$ be an eigenvector of C with eigenvalue $\lambda > 0$, then $P^T v$ is an eigenvector of G also with eigenvalue λ as can be seen from the following simple calculation:

$$GP^T v = P^T P P^T v = P^T C v = \lambda P^T v.$$

Similarly, if $u \in \mathbb{R}^n$ is an eigenvector of G with eigenvalue $\mu > 0$, then Pu is an eigenvector of C with eigenvalue μ. □

One important issue with both PCA and MDS is how to choose/determine k (the intrinsic dimensionality of the point cloud data). Sometimes there is a "large" gap in the eigenvalue spectrum of C or G, respectively, and k is then often chosen as the number of eigenvalues above this gap.

2.1. PCA. As mentioned earlier PCA asks for the k-dimensional subspace of $\mathbb{R}^d$ that approximates the point set P best possible in a least squares sense. Let us discuss this for the case $k = d - 1$ first. In this case we are looking for a unit vector $v \in \mathbb{R}^d$ such that the sum of the squared lengths of the projections $(v^T p_i)v$ is minimized. Formally this can be written as

$$\begin{array}{ll} \min & v^T P P^T v \\ \text{s.t.} & \|v\|^2 = 1. \end{array}$$

From the Lagrange multiplier theorem one derives the following condition for an optimal solution to this optimization problem: $\lambda v = PP^T v = Cv$. That is, an optimal solution is the subspace orthogonal to an eigenvector of the covariance matrix C and the value of the optimization problem at an optimal solution is $v^T PP^T v = v^T C v = \lambda \|v\|^2 = \lambda$. Hence we are looking for an eigenvector associated to the smallest eigenvalue of C and the optimal solution is spanned by all eigenvectors of the covariance matrix C except the one corresponding to the smallest eigenvalue.

Let $\lambda_1 \geq \ldots \geq \lambda_d \geq 0$ be the eigenvalues of C, $v_1, \ldots, v_d \in \mathbb{R}^d$ a corresponding orthonormal eigenbasis and $P_k = \sum_{i=1}^k v_i v_i^T, k = 1, \ldots, d$, the projector on the k'th invariant eigenspace, i.e., the eigenspace spanned by the first k eigenvectors. Iteratively it follows that the best approximating k-dimensional subspace of $\mathbb{R}^d$ in a least square sense is spanned by $v_1, \ldots, v_k$. The k'th order PCA is then given as the following transformation:

$$p_i \mapsto P_k p_i = p_i - (\mathbb{I} - P_k)p_i.$$

In a way $P_k p_i$ is seen as the signal conveyed with the point p_i and $(\mathbb{I} - P_k)p_i$ is seen as noise.

2.2. MDS. Multi-dimensional scaling is aiming for a k-dimensional embedding of P that preserves the pairwise inner products $p_i^T p_j$ as well as possible in a least squares sense[3]. Note that all inner products are stored as entries in the Gram matrix G. Let $\mu_1 \geq \ldots \geq \mu_n \geq 0$ be the eigenvalues of G, let $u_1, \ldots, u_n \in \mathbb{R}^n$ be a corresponding orthonormal eigenbasis and let $Q_k = \sum_{i=1}^{k} u_i u_i^T$, for $k = 1, \ldots, n$, be the projector on the k'th invariant eigenspace. We have the following observation:

OBSERVATION 2. *The matrix $Q_k G$ is the best rank k approximation of the Gram matrix G in the sense that*

$$\|Q_k G - G\|_2 = \mathit{argmin}_{Q:\,(n\times n)\text{-}\mathit{matrix\ of\ rank\ }k} \|QG - G\|_2.$$

The matrix $Q_k G$ can also be interpreted as a matrix of inner products. To see this we use (a) the projector property $Q_k^2 = Q_k$, (b) symmetry $Q_k^T = Q_k$, and (c) the commutator property $Q_k G = G Q_k$, and get

$$Q_k G = Q_k^2 G = Q_k G Q_k = Q_k P^T P Q_k = Q_k^T P^T P Q_k = (PQ_k)^T P Q_k,$$

which shows that $Q_k G$ is the matrix of inner products of the columns of $PQ_k = (Q_k P^T)^T$. Here the $(n \times d)$-matrix $Q_k P^T$ is the projection of the rows of P onto the space spanned by $u_1, \ldots, u_n$. The k'th order MDS maps the point p_i to the i'th column $Q_k P^T$, i.e., the i'th column of PQ_k. This column is uniquely specified by its coefficients $\alpha_1^i, \ldots, \alpha_k^i$ in the orthonormal basis $u_1, \ldots, u_k$. Representing the points p_i by $(\alpha_1^i, \ldots, \alpha_k^i)$ gives the thought for least squares optimal k-dimensional embedding of the point set P.

3. Localization.

3.1. Neighborhood criteria. In using PCA and MDS, feature preserving data quantization and visualization can be enhanced by taking only local relations among all the data points into account. Localization the relations means choosing neighborhoods for each data point, i.e., building a (in general directed) neighborhood graph on the data points. The right choice of neighborhood is crucial for the localized version of PCA and MDS to work properly. Commonly used methods to define the neighborhoods are:

(1) κ nearest neighbors: connect every p_i to its κ nearest neighbors (in terms of Euclidean distance) in P.
(2) symmetric κ nearest neighbors: connect p_i to its κ nearest neighbors and all points in this neighborhood to each other.
(3) fixed neighborhood: given $\varepsilon > 0$, connect every p_i to all points in P that have distance less than ε to p_i.
(4) relative neighborhood: given $\rho > 1$, connect every p_i to all neighbors at distance ρ times the distance of p_i to its nearest neighbor.

[3]Observe that completely preserving the inner products allows us to recover P up to a rotation, i.e., completely preserving the pairwise inner products also preserves the pairwise distances $\|p_i - p_j\|$.

An important observation is that (1) and (2), i.e., κ nearest neighbors and symmetric κ nearest neighbors, respectively, do not automatically adapt to the intrinsic dimension of the point cloud data. Intuitively, if the intrinsic dimension is large also κ needs to be large in order to cover a meaningful neighborhood for a data point (we expect this neighborhood to grow exponentially in the intrinsic dimension), whereas if the intrinsic dimension is small, for the same value of κ one already covers data points quite far away. Methods (3) and (4), fixed- and relative neighborhood, both automatically adapt to the intrinsic dimension, but cannot—in contrast to the κ nearest neighbor methods—adapt to non-uniform or anisotropic spacing of the data points. In practice a good choice for the value of the parameter ρ of (4) may be easier to find than for the value of ε in (3).

More neighborhood graphs are discussed by Yang [56] who also provides experimental results.

3.2. Dimension detection using PCA. We have seen above that knowing the local dimension at a data point can guide the right choice of parameter κ when computing the κ nearest neighbors neighborhood. On the other hand, using that given $p \in M$, there exists a small neighborhood of p in which M is close to its tangent space at p, it is appealing to use localized versions of PCA to infer the local intrinsic dimension of M at p from the point cloud data P. With a good neighborhood $N(p) \subset P$ of $p \in P$ one can estimate the intrinsic dimension at p by a localized version of PCA. The localized version uses the local covariance matrix C_p of the points

$$p_i' = (p_i - p) - \frac{1}{n} \sum_{p_i \in N(p)} (p_i - p) \text{ for } p_i \in N(p).$$

Intuitively, if the local dimension at p is k, then we expect a gap in the eigenvalue spectrum of C_p in the sense that k'th largest eigenvalue is much larger than the $(k+1)$'st eigenvalue and the k largest eigenvalues are roughly of the same magnitude. That is, we expect a threshold θ such that

$$\frac{\lambda_j}{\lambda_1} \geq \theta \text{ for } j \leq k \text{ and } \frac{\lambda_j}{\lambda_1} \leq \theta \text{ for } j > k.$$

Indeed, Cheng, Wang and Wu [13] were able to prove the existence of such a threshold θ under the assumption that the data are sampled from a smooth manifold and obey a sampling condition. The sampling condition rules out locally non-uniform or anisotropic spacing of the sample points. Under this assumption fixed- and relative neighborhoods should work. Cheng et. al use the relative neighborhood for their proof. Though their threshold parameter θ depends on parameters of the sampling condition they report good results in practice using a threshold of $\theta = 1/4$.

It is important to remark that when the sampling conditions are not fullfilled or when the size of the neighborhoods are not well-choosen, the

previous method usually leads to unclear and confusing estimations. In particular the dimension estimation may depend on a "scale" (in the previous case the size of the neighborhoods) at which the data is considered: assume that P samples a planar spiral with gaussian noise in the normal direction to the curve. At a "microscopic" scale, P just looks like a finite set of points and its dimension is 0. At a scale of the size of the standard deviation of the noise, P seems to locally sample the ambient space and the localized PCA method will probably estimate M to be 2-dimensional. At a higher, but not too big, scale the localized PCA will provide the right estimation and at large scales, it will again provide a 2-dimensional estimation. Various notions of dimension (q-dimension, capacity dimension, correlation dimension, etc...) have been introduced to define the intrinsic dimension of general shapes (including non smooth shapes and fractal sets). They give rise to algorithmically simple methods that simultaneously provide dimension estimations at different explicit scales allowing the user to select the one which is most relevant for his purpose. An introduction to this subject may be found in [39].

4. Turning non-linear. The linear and global aspects of PCA and MDS make them inefficient when the underlying manifold M is *highly non linear*. Designing non-linear dimensionality reduction methods that lead to good results for non linear smooth manifolds is an active research area that gave rise to a big amount of literature during the last decade. In this section, we quickly present a set of quite famous dimension reduction methods that take advantage of the localization techniques presented in the previous section and that have interesting geometric interpretations. They also have the advantage of leading to easy to implement polynomial time algorithms that prove more efficient with larger data sets than the ones usually involved in iterative or greedy methods (like e.g. the ones involving EM or EM-like algorithms). We also discuss the guarantees provided by these methods.

Recall that in the following the considered data sets $P \subset \mathbb{R}^d$ are assumed to be sampled on/around a possibly unknown smooth manifold M of dimension k. The common thread of the few methods presented below is that they all aim to find a projection $\hat{P} \subset \mathbb{R}^k$ of the data set minimizing a quadratic functional $\phi(\hat{P})$ that intends to preserve (local) neighborhood information between the sample points.

4.1. Maximum variance unfolding (MVU). PCA and MDS perform poorly when data points are not close to an affine subspace, i.e., they are both based on an inherent linearity assumption. Especially, both methods fail when the data points are close to a "curled up" linear space—the most famous example is the so called Swiss roll data set, points sampled densely from a curled up planar rectangle in $\mathbb{R}^3$. The idea behind *maximum variance unfolding (MVU)*, introduced by Weinberger and Saul [52, 55, 54], is to unfold the data, i.e., to transform the data set to a locally isometric

data set, that is closer to an affine subspace. The unfolding aims at maximizing the distance between non-neighboring points (after some choice of neighborhood) while preserving the distances between neighboring points.

Technically MVU proceeds as follows: let $D = (d_{ij} = \|p\_p_j\|^2)$ be the symmetric $(n \times n)$-matrix of pairwise distances. Choose a suited neighborhood for each point in P (Weinberger and Saul choose the symmetric κ-nearest neighbors) and let the indicator variable n_{ij} be 1 if either p_i is in the neighborhood of p_j or p_j is in the neighborhood of p_i, and 0 otherwise. From D an *unfolding*, a positive semi-definite $(n \times n)$-matrix $K = (k_{ij})$ (interpreted as the Gram matrix of the unfolded point set) is computed through the following semi-definite program (SDP)

Maximize the trace of K subject to

(1) K is positive semi-definite

(2) $\sum_{i,j=1}^{n} k_{ij} = 0$

(3) $k_{ii} - 2k_{ij} + k_{jj} = d_{ij}$ for all (i, j) with $n_{ij} = 1$.

From K a lower dimensional embedding can be computed as described for MDS.

4.2. Locally linear embedding (LLE). LLE is a method introduced in [45, 46] that intends to take into account the local linearity of the underlying manifold M to perform the reduction of dimension. In a first step, LLE discards pairwise distances between widely separated points by building a neighborhood graph G (see Section 3). The goal of this first step is to connect only close points of P so that the neighbors of each vertex p_i in G are contained in a small neighborhood of p_i which is close to the tangent space of the underlying manifold M at p_i. To take this local linearity into account, LLE computes for each vertex p_i of the graph its best approximation as a linear combination of its neighbors. More precisely, one computes a sparse matrix of weights $W_{i,j}$ that minimize the quadratic error

$$\varepsilon(W) = \sum_{i=1}^{n} \|p_i - \sum_{j \in N(p_i)} W_{i,j} p_j\|^2$$

where $N(p_i)$ is the set of the vertices that are connected to p_i in G. This is a simple least square problem. Solving it with the additional constraint

$$\forall i, \quad \sum_{j \in Ngb(p_i)} W_{i,j} = 1$$

makes the weights invariant to rescaling, rotations and translations of the data (the weights thus characterize intrinsic properties of the data). The weights matrix is then used to perform the dimensionality reduction: given $k < d$, the points p_i are mapped to the points $\hat{p}_i \in \mathbb{R}^k$ that minimize the quadratic function

$$\Phi(\hat{p}_i) = \sum_i \|\hat{p}_i - \sum_j W_{i,j}\hat{p}_j\|^2$$

This quadratic minimization problem classically reduces to solving a sparse $n \times n$ eigenvalue problem. As for MDS, the LLE algorithm projects the data in a low dimensional space, no matter what the mapping is. To provide satisfactory result, the data have to be sufficiently dense to insure that the neighbors of a given point provide a good approximation of the tangent space of M. Moreover, even if the data are dense enough, the choice of the neighbors may also be awkward: choosing a too small or too large neighborhood may lead to very bad estimates of the tangent space.

4.3. ISOMAP. ISOMAP is a version of MDS introduced in [48, 21], where the matrix of inner products or Euclidean distances, respectively, is replaced by the matrix of the geodesic distances between data points on M. In a first step, ISOMAP builds a neighborhood graph such that the distances between points of P in the graph are close to the geodesic distances on M. Once the geodesic distance matrix has been built, ISOMAP proceeds like classical MDS to project P in $\mathbb{R}^k$.
One of the advantage of ISOMAP is that it provides convergence guarantees. First, it can be proven that if the data are sufficiently densely sampled on M, the distance on the neighbor graph is close to the one on M [20, 44, 26]. Nevertheless, in practice robust estimation of geodesic distances on a manifold is an awkward problem that requires rather restrictive assumptions on the sampling. Second, since the MDS step in the ISOMAP algorithm intends to preserve the geodesic distances between points, it provides a correct embedding if M is isometric to a convex open set of $\mathbb{R}^k$. The convexity constraint comes from the following remark: if M is an open subset of $\mathbb{R}^k$ which is not convex, then there exist a pair of points that cannot be joined by a straight line contained in M. As a consequence, their geodesic distance cannot be equal to the Euclidean distance. It appears that ISOMAP is not well-suited to deal with data on manifolds M that do not fulfill this hypothesis. Nevertheless some variants (conformal ISOMAP [21]) have been proposed to overcome this issue. Note also that ISOMAP is a non local method since all geodesic distances between pairs of points are taken into account. As a consequence ISOMAP involves a non-sparse eigenvalue problem which is a main drawback of this method. To partly overcome this difficulty some variant of the algorithm using landmarks have been proposed in [21].

4.4. Laplacian eigenmaps. This method introduced in [4, 3] follows the following general scheme: first a weighted graph G with weights $W_{i,j}$ is built from the data. Here the weights measure closeness between the points: intuitively the bigger $W_{i,j}$ is, the closer p_i and p_j are. A classical choice for

the weights is given by the Gaussian kernel $W_{i,j} = \exp(-\frac{\|p_i - p_j\|^2}{4\sigma})$, where σ is a user-defined parameter[4]. Second the graph G is embedded into $\mathbb{R}^k$ in such a way that the close connected points stay as close as possible. More precisely the points p_i are mapped to the points $\hat{p}_i \in \mathbb{R}^k$ that minimize

$$\phi(\hat{P}) = \sum_{i,j} \|\hat{p}_i - \hat{p}_j\|^2 W_{i,j}.$$

There is an interesting and fundamental analogy between this discrete minimization problem on the graph G and a continuous minimization problem on M. Indeed, it can be seen that minimizing ϕ on the functions defined on the vertices of G corresponds (in a discretized version) to minimizing $\int_M \|\nabla f\|^2$ on the space of functions f defined on M with L^2 norm $\|f\|_{L^2}^2 = \int_M \|f\|^2 = 1$. From the Stokes formula, this integral is equal to $\int_M \mathcal{L}(f)f$, where $\mathcal{L}$ is the Laplace-Beltrami operator on M and its minimum is realized for eigenfunctions of $\mathcal{L}$. Similarly the minimization problem on G boils down to a general eigenvector problem involving the Laplacian of the graph. Indeed the Laplace operator on G is the matrix $L = D - W$, where D is the diagonal matrix $D_{i,i} = \sum_j W_{i,j}$. It can be seen as an operator acting on the functions f defined on the vertices of G by subtracting from $f(p_i)$ the weighted mean value of f on the neighbors of p_i. By a classical computation, one can see that $\phi(\hat{P}) = tr(\hat{P}^T L \hat{P})$, where $\hat{P}$ is the $n \times k$ matrix with i-th row given by the coordinates of $\hat{p}_i$. It follows that, given $k > 0$, the minimum of ϕ is deduced from the computation of the $k+1$ smallest eigenvalues of the equation $Ly = \lambda Dy$ (the smallest one corresponding to the eigenvalue 0 has to be removed). The analogy between the discrete and continuous setting extends to the choice of the weights of G: choosing $W_{i,j} = \exp(-\frac{\|p_i - p_j\|^2}{4\sigma})$, where σ is a user-defined parameter, allows to interpret the weights as a discretization of the heat kernel on M [4]. From the side of the guarantees, the Laplacian eigenmaps only involve intrinsic properties of G so they are robust to isometric perturbations of the data. Moreover, the relationship with the Laplacian operator on M provides a framework leading to convergence results of L to the Laplace operator on M [3].

4.5. Hessian eigenmaps (HLLE). ISOMAP provides guarantees when the unknown manifold M is isometric to a convex open subset of $\mathbb{R}^k$. Although the hypothesis of being isometric to an open subset of $\mathbb{R}^k$ seems to be rather reasonable in several practical applications, the convexity hypothesis appears to be often too restrictive. HLLE is a method introduced in [24] intending to overcome this convexity constraint. The motivation of HLLE comes from a rather elementary result stating that if M is isometric to a connected open subset of $\mathbb{R}^k$ then the null-space of the operator defined on the space of $\mathcal{C}^2$-functions on M by

[4]To obtain a sparse matrix W the values of $W_{i,j}$ that are smaller than some fixed small threshold are usually set to 0.

$$\mathcal{H} : f \rightarrow \int_M \|Hessf(m)\|^2 dm$$

where $Hessf$ is the Hessian of f, is a $(k+1)$-dimensional space spanned by the constant functions and the "isometric coordinates" of M. More precisely, if there exists an open set U in $\mathbb{R}^k$ and an isometric embedding $\phi : M \rightarrow U$ then it can be proven that the constant functions and the functions $\phi_1, \cdots \phi_k$, where ϕ_i is the i-th coordinate of the map ϕ, are contained in the null-space of $\mathcal{H}$. Moreover, the constant functions span one dimension of this null-space and the k functions ϕ_i span the k other dimensions. It is thus appealing to estimate this null space in order to recover these isometric coordinates to map M isometrically on an open subset of $\mathbb{R}^k$. To do this the algorithm follows the same scheme as LLE and the estimation of the null-space of $\mathcal{H}$ reduces to an eigenvalue computation of a sparse $n \times n$ matrix. As a consequence HLLE allows to process dimensionality reduction for a larger class of manifolds M than ISOMAP. The quality of the reduction is obviously closely related to the quality of the approximation of the kernel of the operator $\mathcal{H}$. Nevertheless, it is important to notice that the algorithm involves the estimation of second order differential quantities for the computation of the Hessian while LLE requires only first order ones to approximate the tangent space of M. To be done efficiently this usually needs a very dense sampling of M. At last, note that HLLE is the same as Laplacian Eigenmaps where the Laplacian operator has been replaced by $\mathcal{H}$.

4.6. Diffusion maps. Diffusion maps [15] provide a method for dimensionality reduction based upon Markov random walks on a weighted graph G reflecting the local geometry of P . The graph G is built in a similar way as for Laplacian Eigenmaps: the larger is the weight of an edge, the "closer" are its endpoints. In particular G can be built using the discretization of the heat kernel on M (see Section 4.4). From the weights matrix W one constructs a Markov transition matrix Π by normalizing the rows of W

$$\Pi_{i,j} = \frac{W_{i,j}}{d(p_i)} \quad \text{where} \quad d(p_i) = \sum_k W_{i,k} \quad \text{is the degree of the vertex} \quad p_i$$

$\Pi_{i,j}$ can be interpreted as the probability of transition from p_i to p_j in one time step. The term $\Pi_t(i,j)$ of the successive powers Π^t of Π represent the probability $\Pi_t(p_i, p_j)$ of going from p_i to p_j in t steps. The matrix Π can be seen as an operator acting on the probability distributions supported on the vertices of G. It admits an invariant distribution ϕ_0 defined by $\phi_0(p_i) = \frac{d(p_i)}{\sum_j d(p_j)}$. The idea of diffusion maps is thus to define a metric between the points of P which is such that at a given $t > 0$ two points p_i and p_j are close if the conditional distributions of probability $\Pi_t(p_i, .)$ and $\Pi_t(p_j, .)$ are close. The choice of a weighted L^2 metric between the

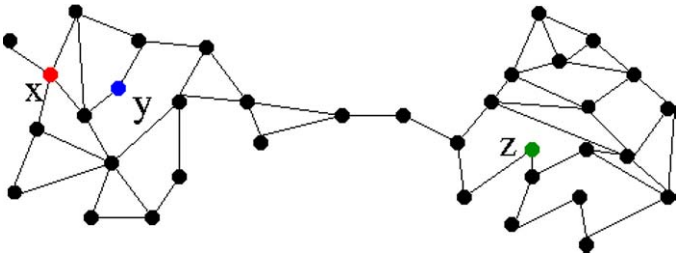


FIG. 1. *An example of a graph G (the weights are given by the heat kernel approximation, see text) with points that are close or far to each other with respect to the diffusion metric: the points x and y are close to each other while the points x and z are far away because G is "pinched" between the two parts containing x and z. So there are few paths connecting x to z.*

conditional distributions allows to define a *diffusion metric* between the points of P

$$D_t^2(p_i, p_j) = \sum_k \frac{(\Pi_t(p_i, p_k) - \Pi_t(p_j, p_k))^2}{\phi_0(p_k)}$$

which is closely related to the spectral properties of the random walk on G given by Π. Intuitively, two points p_i and p_j are close if there are many paths connecting them in G as illustrated on Fig. 1. Note that the parameter t representing the "duration" of the diffusion process may be interpreted as a scale parameter in the analysis. Given k and $t > 0$, the *diffusion map* provides a parameterization and a projection of the data set which performs a dimensionality reduction that minimizes the distortion between the Euclidean distance in $\mathbb{R}^k$ and the diffusion distance D_t. The diffusion map is obtained from the eigenvectors of the transition matrix Π and the eigenvalues to the power t of the transition matrix. The diffusion maps framework reveals deep connections with other areas (such as spectral clustering, spectral analysis on manifolds,...) that open many questions and make it an active research area. For a more detailed presentation of diffusion maps and its further developments the reader is referred to [15, 16, 37].

5. Applications in structural biology: the folding problem. In this section, we first recall the intrinsic difficulty of folding proteins on a computer –Section 5.1, and bridge the gap between folding and dimensionality reduction –Section 5.2. We then proceed with a detailed account of the bio-physical context by discussing the question of cooperative motions within a protein –Section 5.3, and make the connexion to Morse theory and singularity theory along the way. Finally, we review techniques to derive meaningful so-called reaction coordinates –Section 5.4.

5.1. Folding: from experiments to modeling. Anfinsen was awarded the 1972 Nobel prize in chemistry *for his work on ribonuclease, especially concerning the connection between the amino acid sequence and the biologically active conformation*[5]. Since then, Anfinsen's dogma states that for (small globular) proteins, the sequence of amino-acids contains the information that allows the protein to fold i.e. to adopts its (essentially unique) native structure, or phrased differently, the 3d structure that accounts for its function. At room temperature, the folding of a protein typically requires from millisecond to seconds, while the time-scale of the finest (Newtonian) physical phenomena involved is the femtosecond.

When compared to femtoseconds, folding times are rather slow, which points towards a process more complex than a mere descent towards a minimum of energy. On the other hand, such folding times are definitely too fast to be compatible with a uniform exploration of an exponential number conformations[6]. This observation is known as Levinthal's paradox [90], and scales the difficulty of folding from a computational perspective.

5.2. Energy landscapes and dimensionality reduction.

5.2.1. Potential and free energy landscape. Consider a system consisting of a protein and the surrounding solvent, for a total of n atoms. Each atom is described by 3 parameters for the position and 3 for its velocity (momentum). In the following, depending on the context, we shall be interested in a parameter space of dimension $d = 3n$ (positions) or $d = 6n$ (positions+velocities), the latter being called the *phase* space. As the system is invariant upon rigid motions, one could work with $d - 6$ degrees of freedom, but we skip this subtlety in the following. From this d-dimensional parameter space, one defines the *the energy landscape* [99], i.e. the d dimensional manifold obtained by associating to each conformation of the system an energy (potential energy or free energy[7]). Since the water molecules are critical to model appropriately the electrostatic interactions, n typically lies in the range 10^4 to 10^5 for a system consisting of a protein and its aqueous environment.

5.2.2. Enthalpy-entropy compensation, energy funnel, ruggedness and frustration.

Energy funnels. Folding may be seen as the process driving a heterogeneous ensemble of conformations populating the unfolded state to a homogeneous ensemble of conformations populating the folded or native state. To intuitively capture one major subtlety of folding, it is instructive to examine the variation of the enthalpy H and entropy S of the system

[5]See `http://nobelprize.org/nobel_prizes/chemistry/laureates/1972/index.html`

[6]Recall that the side-chains of the amino-acids take conformations within finite sets –the so-called rotamers [86, 73], whence a priori an exponential number of conformations.

[7]As will be shown with Eq. (2), a free energy landscape is obtained from the potential energy landscape by projecting onto selected coordinates.

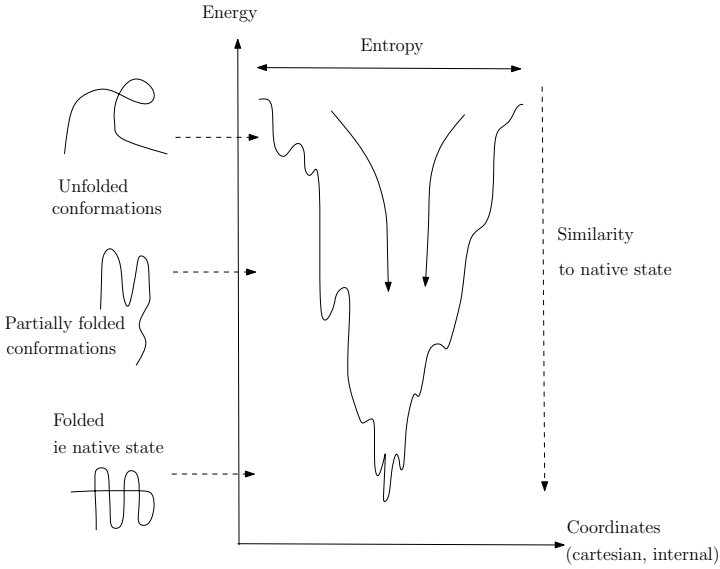


FIG. 2. *Folding funnel. The variability of conformations encodes the entropy of the system, while its energy level encodes the proximity to nativeness. Adapted from [67].*

protein+solvent. While the protein folds, more native contacts between atoms get formed, whence an enthalpy decrease. On the other hand, two phenomena account for an entropic drop down: first, the conformational variability of the protein decreases; second, the structure of the solvent around the protein changes. This latter re-organization, known as the hydrophobic effect, corresponds to the fact that water molecules line-up along the hydrophobic wall formed by the molecular surface of the protein. Overall, the variations of the enthalpy and entropy almost cancel out, resulting in a small variation of the free energy $G = H - TS$ of the system. This phenomenon is known as the *enthalpy-entropy compensation* [75], and can be illustrated using energy landscapes, as seen from Fig. 2. On this figure, the vertical axis features the free energy, and the horizontal one the entropy: while the folding process progresses, the free energy (slightly) decreases and the landscape becomes narrower—the entropy decreases. Such a landscape is generally called a *folding funnel [67].*

While the previous discussion provides a thermodynamic overview of the folding process, Levinthal's paradox deals with a kinetic problem—*how come the folding process is so fast?* Travelling down the folding funnel[8] provides an intuitive explanation: the protein is driven towards the minimum of energy corresponding to the native state by a steep gradient along the

[8]The fact that the kinetic pathway follows the thermodynamic one is non trivial, and in general unwarranted, see [75, Chapter 19].

energy surface. This intuitive simplified view, however, must be amended in several directions.

Ruggedness and frustration. Two important concepts which help to describe landscapes are *ruggedness* or *roughness* and *frustration*. Ruggedness refers to the presence of local minima, which in a folding process may correspond to partially folded states. Frustration refers to the presence of several equally deep minima separated by significant barriers, which may prevent the system from reaching the deepest one. The fact that most proteins seem to have a single native state[9] seems to advocate a minimal frustration principle. Yet, even for non frustrated landscapes, several levels of *ruggedness* may exist. In particular, on the easy side of the spectrum, one finds proteins folding with a two-states kinetics, i.e. without any intermediates [85].

Ruggedness / frustration may actually come from two sources, namely from the interaction energy between atoms of the protein, and/or from the conformational entropy [68]. The enthalpic frustration comes from local minima of the interaction potential energy. For the entropic frustration, observe that the folding process is accompanied by a loss of conformational entropy (of the protein). If this loss is heterogeneous and larger than the energetic heterogeneity, the corresponding free energetic landscape becomes frustrated.

5.2.3. Cooperativity and correlated motions. Another concept related to minimally frustrated folding funnels is that of *cooperative* motions between atoms. Cooperativity stipulates that when one atom is moving, atoms nearby must move in a coherent fashion. This is rather intuitive for condensed states where local forces (repulsion forces as atoms cannot inter-penetrate, hydrogen bonding) are prominent. At a more global scale, cooperation is likely to also be important, e.g. due to electrostatic interactions. From a technical point of view, simple illustrations of correlated motions are provided by normal modes studies[10], as well as correlations between positional fluctuations[11].

Having mentioned correlated motions of atoms, the fact that dimensionality reduction techniques play a key role in modeling folding (and more generally the behavior of macro-molecular systems) is expected. First, the

[9] As opposed to many polymers which exist under a number of energetically equivalent inter-convertible states.

[10] Assuming the system is at a minimum of its potential energy V, the dominant term in the Taylor series expansion of V is the quadratic one. Diagonalizing the corresponding quadratic form yields the so-called normal modes, whose associated eigenvectors are collective coordinates. See for example [97].

[11] Given a molecular dynamics simulation, one may investigate the correlations between the atomic fluctuations —with respect to a reference conformation. Both PCA and MDS have been used for this problem: in [78, 59], the average covariance matrix of the positional fluctuations is resorted to, while [83] computes the average Gram matrix. See also [89] for a characterization of pairwise atomic correlations based of Pearson's coefficients and relatives.

d degrees of freedom are certainly not equivalent, as different time-scales are clearly involved: from small amplitude - high frequency vibrations apart from chemical bonds, to large amplitude - slow frequency deformations of the protein. Second, the constraints inherent to the large amplitude motions are such that one expects the *effective* parameters to lie on some lower-dimensional manifold representing the *effective* energy landscape, that is the one accounting for transitions.

5.3. Bio-physics: Pre-requisites.

5.3.1. Molecular dynamics simulations. The simulation data we shall be concerned with are essentially molecular dynamics (MD) data. (The reader is referred to [77] for alternate simulation methods, such as Monte Carlo simulations or Langevin dynamics.) A MD simulation is a deterministic process which evolves a system according to the Newtonian laws of motion. Central in the process is the force field associated to the system, or equivalently the potential energy stemming from the interactions between atoms. A typical potential energy involves bonded terms (energies associated to covalent bonds), and non bonded terms (Van der Waals interactions and electrostatic interactions). From the potential energy V associated to two atoms, one derives an associated force. Given these forces, together with the positions and momenta of the atoms, one determines the configuration of the system at time $t + \Delta t$. Practically, Δt is of the order of the femtosecond, so that in retaining one conformation every 10, long simulations (beyond the nanosecond) result in a number of conformations $> 100,000$.

5.3.2. Models, potential energy landscapes and their ruggedness. As exploring exhaustively the energy landscape of large atomic models is not possible, a number of coarse models mimicking the properties of all atoms models have been developed. We may cite the united residues model [74]; the BLN model [63] which features three types of beads only (hydrophilic, hydrophobic, neutral); the 20 colors beads model [71], which accommodates anisotropic interactions between residues so as to maximize packing of side chains.

Such coarse models deserve a comment about the ruggedness of potential energy landscapes. Ruggedness and frustration are indeed clearly related to the complexity of the force field governing the system, since non local interactions between atoms are likely to yield local minima of the landscape —cf the G$\bar{o}$ models thereafter. On the other hand, non local interactions are likely to help the protein to overcome local energy barriers (to escape the local minima of the rugged landscape) due to solvent collisions, non-native contacts, etc. See for example [92].

Having mentioned energy landscapes and MD simulations, a crucial remark is in order. Following the gradient vector field of the energy on the potential energy surface amounts to a mere energy minimization. But

MD simulations are more powerful, since a system evolved by a MD can cross energy barriers thanks to its kinetic energy[12]. Another way to cross barriers is to resort to a Monte Carlo simulation [77].

5.3.3. Morse theory and singularity theory. As outlined above, the properties of a system are described by its energy landscape. To investigate transitions of our macro-molecular system, the topographical features of the landscape i.e. its minima, maxima, and passes are of utmost importance [99]. These features are best described in terms of Morse theory [91] as well as singularity theory [23], which in our setting amounts to studying the gradient vector of the energy function on the manifold.

Following classical terminology, a *critical* point of a differentiable function is a point where the gradient of the function vanishes, and the function is called a *Morse* function if its critical points are isolated and non-degenerate. For a critical point p of such a function, the stable (unstable) manifold $W^s(p)$ ($W^u(p)$) is the union of all integral curves associated to the gradient of the function, and respectively ending (originating) at p. Locally about a critical point of index i (the Hessian has i negative eigenvalues), the (un-)stable manifold is a topological disk of dimension i ($d-i$). The stable and unstable manifolds are also called the *separatrices*, as they partition the manifold into integral curves having the same origin and endpoint. In a more prosaic language, they are also called watersheds, by analogy with water drainage. In particular, under mild non degeneracy assumptions of the energy landscape, a transition between two adjacent minima is expected to correspond to the stable manifold of index one saddle joining the minima –a result known as the Murrell-Laidler theorem in bio-physics [99].

If Morse theory provides a powerful framework to describe energy landscapes, the pieces of information provided should be mitigated by the relative energies associated to critical points of various indices. As already noticed at the end of Section 5.3.2, the thermal energy of the system indeed allows barrier crossing.

5.3.4. Free energy landscapes and reaction coordinates. In classical chemistry, a chemical systems moves from one minimum of energy to another following the minimum energy path, which, as just discussed is expected to go through index one saddles and intermediate minima. For complex systems such as a protein in its aqueous solution, things are more involved [72, 65, 69]. The different parameters have different relaxation times: fast parameters are those describing the solvent, as well as the variables accounting for the fast vibrations apart from covalent bonds of the protein; slow ones account for the large amplitude motions of the protein. Because the system equilibrates faster for some coordinates than others,

[12]If the internal (potential+kinetics) energy remains constant along the MD simulation, the system is Hamiltonian, and a large number of mathematical results apply [94]. We shall get back on this issue in the outlook.

we may partition the parameters as $x = (q, s)$. Denote $V(x)$ the potential energy of the system. By focusing on q and averaging out the other parameters, one defines the free energy landscape, which is the kinetically relevant one, by:

$$W(q) = -kT \ln \int \exp\left[-\frac{V(q,s)}{kT} \right] ds. \tag{2}$$

Coordinates q which provide kinetically relevant informations on transitions are called reactions coordinates. Finding such coordinates is challenging, even on simple systems. We illustrate these difficulties with a two dimensional system corresponding to a two states folding protein, whose unfolded and folded states are respectively denoted A and B. If q is the reaction coordinate sought, obvious requirements are (i) q takes different values q_A and q_B for these states, and (ii) q is such that the free energy W has a maximum at some value q^* in-between q_A and q_B. When these conditions are met, q is called an *order parameter*. If q provides in addition informations about the kinetics of the transitions, it is called a *reaction coordinate*. As illustrated on Fig. 3(a,b), these are different notions. In particular, Fig. 3(b) features a parameter q which is a good order parameter but not a reaction coordinate. For example, the dashed trajectory passes through q^* but does not correspond to a transition. In the ideal setting, for a reaction coordinate, the unstable manifold of the index one saddle joining the two minima separates the points which are committed to one state or the other, and thus determines the so-called Transition State Ensemble (TSE).

Practically, dealing with reaction coordinates poses several problems. First, for a system such as a protein and its solvent, one does not know a priori which variables are the slow ones. This issue is further discussed at the end of Section 5.4.3. Second, if there is not a unique coordinate which is *slower* than the remaining ones, a multi-dimensional analysis must be carried out. Third, computing a free energy profile from Eq. (2) requires the coordinates over which the integration is performed to be equilibrated.

5.3.5. Folding probability p_{fold}. To probe the relevance of a parameter as a reaction coordinate, one resorts to the *committor* probability, i.e. the probability of being committed to a given state [72]. More precisely, for any state x in the system, this is the probability of arriving say at B before before arriving at A within some time t_s. If the potential energy depends on positions and momenta, averaging is understood w.r.t. momenta. By studying this probability along a given path, one locates points near the TSE, since such points are equally committed to both states. Denote Dirac's delta function δ, and let $< z >_E$ the average of quantity z over an ensemble E. To probe the interest of an order parameter as a reaction coordinate, one studies the distribution of the committor probability at $q = q^*$, that is

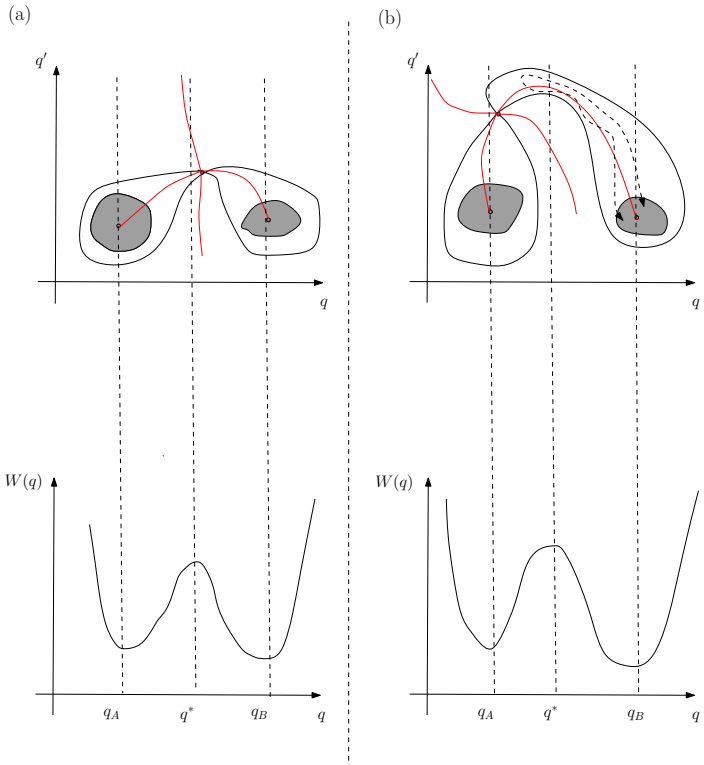


FIG. 3. *Potential energy landscape, with separatrices of the saddle in red: an order parameter may not be a good reaction coordinate. Adapted from [65].*

$$P(p_B) =< \delta[p_B(x,t_s) - p_B] >_{q^*},\ p_B \in [0,1].$$

For a good reaction coordinate, one expects $P(p_B)$ to be sharply peaked at $p_B = 1/2$. This is the case on Fig. 4(a), but not on Fig. 4(b) where $P(p_B)$ is bimodal, meaning that the orthogonal coordinates are such that commitment to the two states occurs. The reader is referred to [72, 66, 65] for example physical systems featuring various committor's distributions.

The notion of transition state is also closely related to that of transition path [82, 64]. Define a transition path TP as a path in phase space that exits a region about the unfolded state, and reaches a region about the folded state. A collection of transition paths determines a conditional phase space density $p(x \mid TP)$, and one has

$$p(TP \mid x) = \frac{p(x \mid TP)p(TP)}{p_{eq}(x)}, \tag{3}$$

with $p_{eq}(x)$ the equilibrium probability of state x and $p(TP)$ the fraction of time spent on transition paths. Transition states are naturally defined

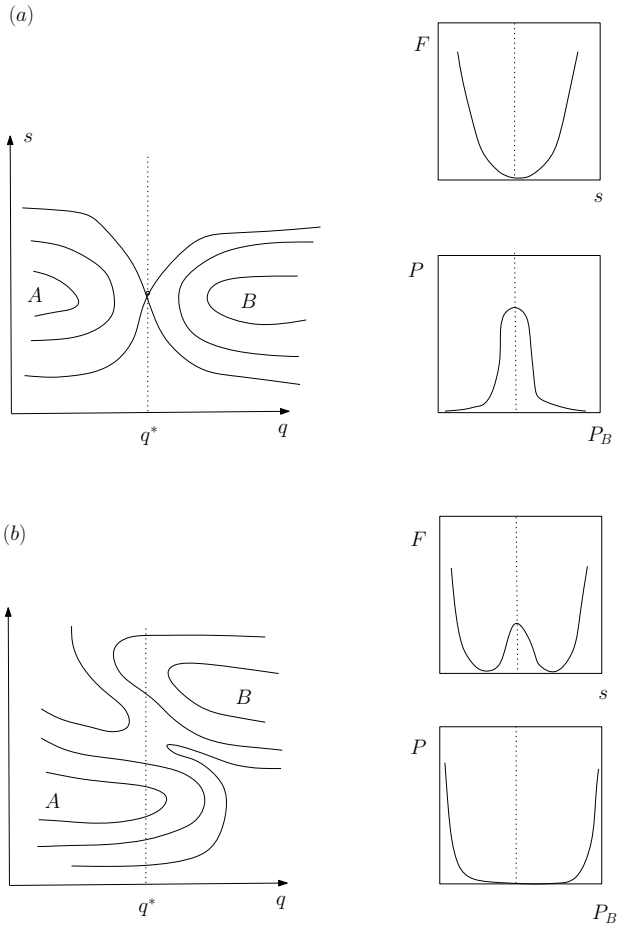


FIG. 4. *Probing a reaction coordinate by computing the committor probabilities p_B. Adapted from [66].*

as points maximizing $p(TP \mid x)$. Moreover, denoting $\overline{x}$ a point with same position and reversed momentum, and $p_A(x)$ the probability of reaching state A before state B from x, it can be shown [82] that

$$p(TP \mid x) = p_A(\overline{x})p_B(x) + p_A(x)p_B(\overline{x}). \tag{4}$$

An important property of this equation is that one can project x onto a lower dimensional space –see Section 5.4. Denoting $r = r(x)$ such a coordinate, it can be shown [82] that

$$p(TP \mid r) = \frac{p(r \mid TP)p(TP)}{p_{eq}(r)}. \tag{5}$$

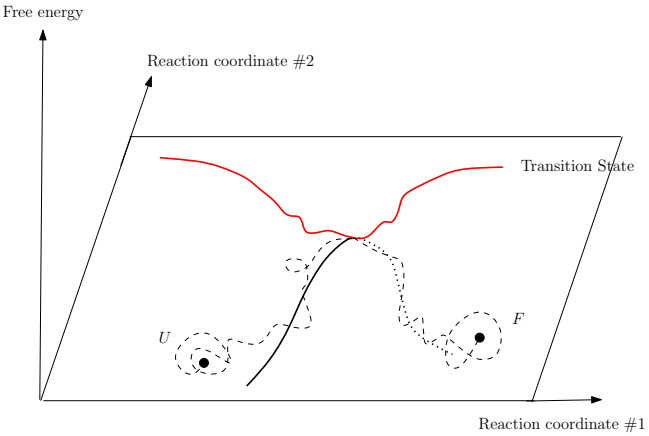


FIG. 5. *Crossing the Transition State on a rugged energy landscape: the system moves from one watershed (state U) to a neighboring watershed (state F) by crossing the energy landscape pass. Adapted from [69].*

Practically, the difficulty with p_{fold} and related quantities are several [69]. First, the concept is bound to simple landscapes corresponding to two states folding processes. Most importantly, estimating p_{fold} requires sampling the TSE, which either requires long simulations –usually out of reach, or some form of importance sampling to favor the rare events corresponding to crossings of the TSE.

5.4. Inferring reaction coordinates. In the following, we review some of the most successful techniques to analyze transitions. We focus on the methodological aspects, and refer the reader to the original papers for a discussion of the insights gained, including connexions with experimental facts. As it can be seen from [69] for example, assessing the relevance of a particular coordinate can be rather controversial.

5.4.1. Reaction coordinates? In order to prove efficient to investigate folding, funnels such as that of Fig. 2 must be made quantitative, that is, one needs to specify what the axis account for. The variables parameterizing the axis are called *reaction coordinates*, and a quantitative energy landscape is displayed on Fig. 5. We now discuss several ways to design such coordinates.

5.4.2. Contacts based analysis. Following the work of G$\bar{o}$ [80], a natural way to tackle Levinthal's paradox consists of introducing a bias in the energy function towards native contacts, i.e. contacts observed in the folded state. More precisely, two residues which are not adjacent along the primary sequence of the protein form a native contact if they are *spatially close* in the protein's native state. Such pairs of residues are associated a favorable interaction energy, while the remaining ones are associated a

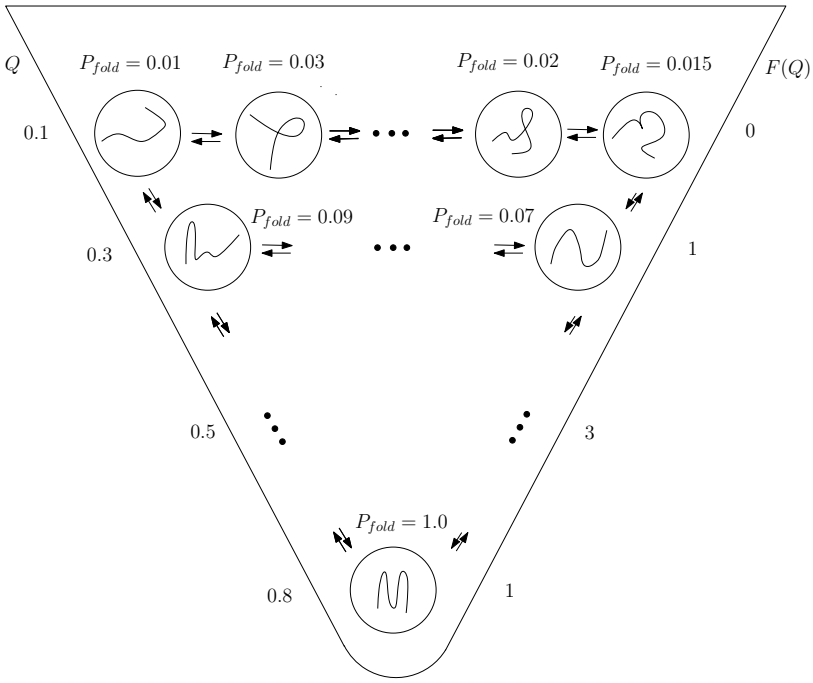


FIG. 6. *Folding process down a folding funnel: fraction of native contacts Q increases, free energy $F(Q)$ crosses a barrier, p_{fold} increases. Adapted From [69].*

repulsive, neutral or less attractive interaction energy. Figure 6 illustrates a folding process down a funnel, described using the fraction of native contacts. On one hand, energy landscapes obtained with G$\overline{o}$ models are generally minimally frustrated. On the other hand, as discussed in Section 5.2, removing non local contacts may impair the folding process. At any rate and regardless of the energy model used, the fraction of native contacts Q can be used as reaction coordinate. Alternative empirical reaction coordinates, also exploiting the resemblance of a particular conformation with the native state, are being used: the radius of gyration (the root mean square distance of the collection of atoms from their center of mass), the effective loop length and the partial contact order [68]. In particular, the latter two coordinates are used in [68] to measure the fraction of conformations that are actually accessible amongst the conformations with the same degree of *nativeness* Q. Such measures are directly related to the entropy of the system along the folding route, and thus allow one to assess the entropic ruggedness of the free energy landscape.

The native contacts can be used in a more elaborate fashion. Following [64], denote Q the matrix such that $Q_{ij} = 1$ if the distance between residues i and j is less than some cutoff (e.g. 12Å), and 0 otherwise. Using a

weight matrix $W = (w_{ij})$, the contact matrix is projected onto a reaction coordinate defined by $r = \sum_{ij} w_{ij} q_{ij}$. Starting from a random initialization of matrix W, the weights are optimized so as to maximize a Gaussian fit of $p(TP \mid r)$ –see Eq. (5). In doing so, one ensures that all reactive configurations are condensed in a single peak.

5.4.3. Dimension reduction based analysis. If one discards the momenta of the points, an important question is to come up with a simplified representation of the $3n$ dimensional energy landscape. Not surprisingly, PCA and MDS have been used for this purpose[13]. A typical illustration is provided by [62], where a PCA analysis of the conformations is first performed. Using the two most informative eigenvectors, an approximation of the landscape termed the *energy envelope* is computed. Fine informations on barriers between watersheds of minima might be lost –the ruggedness observed on a landscape computed from two PCA coordinates is at best questionable, but one expects to retain the overall shape of the watersheds. In [87], a PCA analysis is carried out on the critical points of an energy landscape, rather than on the whole point cloud. This analysis yields new coordinates, which can be plugged into the potential energy function.

One step towards a finer analysis is made in [70], where an adaptation of ISOMAP is used to derive new coordinates. The adaptions w.r.t. the standard ISOMAP algorithm are threefold. First, the computation of the nearest neighbors is done resorting to the least RMSD (lRMSD)[14]. Second, following [21], landmarks are used to alleviate the pairwise geodesic distance calculations. Third, the point cloud is trimmed to get rid of redundancies, which are expected in particular near the minima of potential energy. These conformations are later re-introduced into the low-dimensional embedding, which is important in particular to recover statistical averages. To assess the performance of the dimensionality reduction, a residual variance calculation is performed. For a two states folding protein, the transition state identified from the maximum of the free energy profile $W(x_1)$ associated to the first embedding coordinate x_1 is in full agreement with p_{fold}. (A result also holding for the reaction coordinate Q in this case.) Motivated by the fact that 95% of the running time is devoted to the calculation of nearest neighbors, a further improvement is proposed in [95]. Assume m landmark conformations have been selected. Following the strategy used by the General Positioning System, each conformation (a point $\mathbb{R}^{3n}$) is represented as a m-dimensional point whose coordinates are the lRMSD distances to the m landmarks. In the corresponding m-dimensional space, the $l > k$ nearest neighbors of a point can be computed using the

[13]Notice this analysis is different from the investigation of positional fluctuations mentioned in Section 5.2.

[14]The Root Mean Square Deviation computed once the two structures have been aligned.

Euclidean distance, from which the k nearest ones according to the lRMSD are selected.

To finish up this review, one should mention methods which do not provide a simplified embedding of the landscape, but resort instead to a clustering of the nodes in parameter space [81, 79]. Nodes within the same watershed should belong to the same cluster, from which a Configuration Space Network (CSN) can be built. In some cases, quantitative informations (e.g. free energies) can even be retrieved.

REMARK 5.1. *Having discussed dimensionality reduction techniques, one comment is in order. If one does not know a priory which are the slow variables, integrating Eq. (2) is not possible. This accounts for a three-stage strategy which consists of performing a simulation, performing a dimensionality reduction to infer candidate reaction coordinates, and probing them using* p_{fold}.

5.4.4. Morse theory related analysis. Energy landscapes govern the folding process of proteins, but also the behavior of a number of physical systems such as clusters of atoms, ions or simple molecules [67, 99]. For some of these systems which exhibit a small number of stable crystalline geometries and a large number of amorphous forms, exploring the landscape exhaustively is impossible. Yet, a qualitative analysis can be carried out by focusing on selected critical points. In [88, 60, 63], sequences of triples *minimum - saddle - minimum* are sought, and *super-basins* are built from their concatenation. In a related vein, the relative accessibility of potential energy basins associated to minima is investigated in [61], so as to define the so-called disconnectivity graph (DG). More precisely, two constructions are performed in [61]. The first one, based on the *canonical mapping*, focuses on the relative height of energy barriers, which governs transitions between states, thus encoding the kinetic behavior of the system. The second one, based on the *canonical mapping*, probes the potential energy surface at predefined values of the energy, thus encoding global topological properties of the landscape. Mathematically, constructing either DG is tantamount to tracking the topological changes of the set $V^{-1}(]\infty, v])$ when increasing v. As such changes occur at critical values only [91], the graph built when all critical values are available is called the contour tree[15]. In [61], a discrete set of energies are used to probe the topological changes of the level sets, though.

If one focuses on the relative accessibility of basins, one problem arising is that the DG built does not have any privileged embedding —the vertical axis encodes an energy, but the horizontal one does is meaningless. To complement the topological information by a geometric one, the

[15]Consider the level sets of a Morse function f, and call a connected component of a level set $f^{-1}(h)$ a *contour*. Further contract every contour to a point. The graph encoding the merge/split events between these contours is called the Reeb graph, or the contour tree if the domain is simply connected [76].

following is carried out in [96]: first, similarly to [87], a PCA of critical points is carried out, from which a two-dimensional embedding of these critical points is derived; next, the DG is rendered as a three-dimensional tree, the z coordinate corresponding to the potential energy. Interestingly, such representations convey the (lack of) frustration of BLN models [63], depending on the interaction energy used.

6. Outlook.

Algorithms. Exploring a high-dimensional point cloud with the methods discussed and mentioned raises critical issues which should be kept in mind. First, it is usually assumed that the data points lie on a manifold. But for complex data corresponding e.g. to physical phenomena featuring bifurcations, a stratified complex might actually be the true underlying structure. Even in the manifold case, since the underlying manifold M is unknown, the geometric quantities we aim to preserve have to be estimated from the data set. Coming up with robust estimators poses difficult questions, especially since noisy data (i.e. not exactly sampled on M) has to be accommodated from a practical standpoint. Worse, the sampling conditions insuring that the geometry can be correctly inferred from the data usually depend on some assumptions on M... which is unknown! These questions have been widely studied in computational geometry, in particular for the three dimensional surface reconstruction problem, but remain largely open in a broader setting.

Closely related to the previous questions are those concerning the convergence and theoretical guarantees. As discussed earlier, dimensionality reductions methods are not well suited for all situations. It is thus important to identify the necessary assumptions on M so as to ensure satisfactory results. We have seen in Section 4 that one can answer this question for some of the methods (ISOMAP, HLLE). It is also interesting to have informations on the asymptotic behavior of the considered methods when the samples become denser and denser and converge to M. In this way, Hessian eigenmaps and diffusion maps reveal interesting asymptotic connections with classical operators defined on the underlying manifold M that need to be further explored.

Protein folding. In spite of three decades of intense research, the problem of protein folding is still open. In the context of energy landscapes and dimensionality reduction, a number of further developments are called for.

A variety of (molecular dynamics) simulations are being performed: depending on the system studied (all atoms/coarse, explicit/implicit/no solvent), either the temperature, the pressure or the internal energy of the system are kept constant. For example, if the temperature is held constant using a thermostat —for example the Nose-Hoover, part of the internal energy of the system is dissipated into the thermostat. If the internal energy of the system is conserved, then, the system is Hamiltonian.

For Hamiltonian systems, a large number of mathematical results exist. For example, using the geometrization of Hamiltonian dynamics, a

trajectory of the system corresponds to a geodesic of a suitable Riemannian manifold [94]. This point of view is not really used in recent folding studies, which focus on Morse related analysis of potential and free energy surfaces. The study of the relationship between folding properties inferred from energy landscapes on the one hand, and from Hamiltonian dynamics on the other hand deserves further scrutiny.

Practically, one or two reaction coordinates are usually dealt with, a rather stringent limitation. Methods based on manifold learning are appealing in this perspective, since the dimensionality of the embedding can be estimated. But a critical step for these methods is that of the neighborhood selection. On one hand, the samples are generally processed in a uniform way since the same number of neighbors is used for all points. On the other hand, Morse theory tells us that the local density of samples about a critical point is related to its index. Therefore, a segmentation of the point cloud might be in order before resorting to dimensionality reduction techniques. Doing so might allow one to bridge the gap with Morse theory related methods, whose focus has been on the decomposition of the landscape into basins –as opposed to the design of new coordinates accounting for transitions.

Another key problem is that of stability, in the context of rugged / frustrated landscapes. Ideally, multi-scale analysis of landscapes should be developed, so as to asses what is significant and what is not at a given scale. Topological persistence and more generally tools developed in computational topology might be helpful here. Such analysis might also allow one to spot cascades of minor events in the folding process, such cascades triggering major events –cf phase transitions.

Finally, an improved analysis of landscapes would have another dramatic impact, namely on the simulation processes themselves. Should a finer understanding of cooperative motions be available, steered simulations favoring these coordinates should allow one to move faster along a (rugged) landscape.

Hopefully, a finer geometric and topological analysis of non linearities arising on energy landscapes will help in making simulation able to cope with biologically relevant time scales.

Acknowledgments. F. Cazals wishes to acknoledge Benjamin Bouvier, Ricardo Lima, Marco Pettini and Charles Robert for stimulating discussions.

REFERENCES

[1] P.K. Agarwal, S. Har-Peled, and H. Yu. Embeddings of surfaces, curves, and moving points in euclidean space. In *ACM SoCG*, 2007.

[2] D. Agrafiotus and H. Xu. A self-organizing principle for learning nonlinear manifolds. *PNAS*.

[3] M. Belkin and P. Niyogi. Towards a theoretical foundation for laplacian-based manifold methods. In *COLT 2005*.
[4] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.
[5] M. Belkin and P. Niyogi. Semi-supervised learning on riemannian manifolds. *Machine Learning, Invited, Special Issue on Clustering*, pages 209–234, 2004.
[6] Y. Bengio, M.Monperrus, and H. Larochelle. Nonlocal estimation of manifold structure. *Neural Computation*, 18, 2006.
[7] Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N. Le Roux, and M. Ouimet. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. In *NIPS*, 2004.
[8] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2007.
[9] C.M. Bishop, M. Svensen, and C.K.I. Williams. Gtm: The generative topographic mapping. *Neural Computation*, **10**:215–234, 1998.
[10] M. Brand. Charting a manifold. In *Advances in Neural Information Processing Systems* 15. MIT Press, Cambridge, MA, 2003.
[11] F. Chazal, D. Cohen-Steiner, and A. Lieutier. A sampling theory for compact sets in euclidean space. In *Proceedings of the 22nd ACM Symposium on Computational Geometry*, 2006.
[12] F. Chazal, D. Cohen-Steiner, and Q. Mérigot. Stability of boundary measures. 2007.
[13] Siu-Wing Cheng, Yajun Wang, and Zhuangzhi Wu. Provable dimension detection using principal component analysis. In *Symposium on Computational Geometry*, pp. 208–217, 2005.
[14] B. Christiansen. The shortcomings of nlpca in identifying circulation regimes. *J. Climate*, 18:4814–4823, 2005.
[15] R.R. Coifman, S. Lafon, A. Lee, M. Maggioni, B. Nadler, F. Warner, and S. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proc. of Nat. Acad. Sci.*, **102**:7426–7431, 2005.
[16] R.R. Coifman, S. Lafon, A. Lee, M. Maggioni, B. Nadler, F. Warner, and S. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Multiscale methods. *Proc. of Nat. Acad. Sci.*, **102**:7432–7437, 2005.
[17] J.A. Costa and A.O. Hero. Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *IEEE Trans. on Signal Processing*, **52**(8), 2004.
[18] T.F. Cox and M.A. Cox. *Multidimensional Scaling*. Chapman Hall, 1994.
[19] V. de Silva and G. Carlsson. Topological estimation using witness complexes. In *Eurographics Symposium on Point-BasedGraphics*, ETH, Switzerland, 2004.
[20] V. de Silva, J.C. Langford, and J.B. Tenenbaum. Graph approximations to geodesics on embedded manifolds. 2000.
[21] V. de Silva and J.B. Tenenbaum. Global versus local methods in nonlinear dimensionality reduction. In *Advances in Neural Information Processing Systems* 15. MIT Press, Cambridge, MA, 2003.
[22] M. Dellnitz, M. Hessel von Molo, P. Metzner, R. Preiss, and C. Schutte. Graph algorithms for dynamical systems. In A. Mielke, editor, *Analysis, Modeling and Simulation of Multiscale Problems*. Springer, 2006.
[23] M. Demazure. *Bifurcations and Catastrophes: Geometry of Solutions to Nonlinear Problems*. Springer, 1898.
[24] D. Donoho and C. Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, **100**(10):5591–5596, 2003.
[25] Y. Bengio et al. Learning eigenfunctions links spectral embedding and kernel pca. *Neural compuation*, **16**(10), 2004.

[26] J. Giesen and U. Wagner. Shape dimansion and intrinsic metric from samples of manifolds with high co-dimension. In *Proc. of the 19th Annual symp. Computational Geometry*, pp. 329–337, 2003.

[27] D. Givon, R. Kupferman, and A. Stuart. Extracting macroscopic dymamics. *Nonlinearity*, **17**:R55–R127, 2004.

[28] A. Globerson and S. Roweis. Metric learning by collapsing classes. In *NIPS*, 2005.

[29] Jihun Ham, Daniel D. Lee, Sebastian Mika, and Bernhard Schölkopf. A kernel view of the dimensionality reduction of manifolds. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, p. 47, New York, NY, USA, 2004. ACM.

[30] Gloria Haro, Gregory Randall, and Guillermo Sapiro. Stratification learning: Detecting mixed density and dimensionality in high dimensional point clouds. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pp. 553–560. MIT Press, Cambridge, MA, 2007.

[31] T. Hastie and W. Stuetzle. Principal curves. *J. Amer. Stat. Assoc.*, **84**:502–516, 1989.

[32] Matthias Hein and Markus Maier. Manifold denoising. In *NIPS*, pp. 561–568, 2006.

[33] Matthias Hein and Markus Maier. Manifold denoising. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pp. 561–568. MIT Press, Cambridge, MA, 2007.

[34] I. Horenko, J. Schmidt-Ehrenberg, and C. Schutte. Set-oriented dimension reduction: localizing principal component analysis vie hidden markov models. In *LNBS in Bio-informatics*. 2006.

[35] B. Kégl. Intrinsic dimension estimation using packing numbers. In *Advances in Neural Information Processing Systems* 17. MIT Press, Cambridge, MA, 2002.

[36] R.I. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete structures.

[37] S. Lafon and A.B. Lee. Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning and data set parameterization. *IEEE PAMI*, **28**(9):1393–1403, 2006.

[38] M.C. Law and A.K. Jain. Incremental nonlinear dimensionality reduction by manifold learning. *IEEE Trans. on pattern analysis and machine intelligence*, **28**(3), 2006.

[39] J.A. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction*. Springer, 2007.

[40] Elizaveta Levina and Peter J. Bickel. Maximum likelihood estimation of intrinsic dimension. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*, pp. 777–784. MIT Press, Cambridge, MA, 2005.

[41] Nathan Linial, Eran London, and Yuri Rabinovich. The geometry of graphs and some of its algorithmic applications. In *IEEE Symposium on Foundations of Computer Science*, pp. 577–591, 1994.

[42] J. Mao and A.K. Jain. Artificial neural networks for feature extraction and multivariate data projection. *IEEE Trans. Neural Networks*, **6**(2), 1995.

[43] E. Meerbach, E. Dittmer, I. Horenko, and C. Schutte. Multiscale modelling in molecular dynamics : Biomolecular conformations as metastable states. *Lecture notes in physics*, **703**, 2006.

[44] F. Memoli and G. Sapiro. Distance functions and geodesics on point clouds, 2005.

[45] S.T. Roweis and L.K. Saul. Non linear dimensionality reduction by locally linear embedding. *Science*, **290**:2323–2326, 2000.

[46] S.T. Roweis and L.K. Saul. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, **4**:119–155, 2003.

[47] J.B. Tenenbaum and V. de Silva. Sparse multi-dimensional scaling using landmark points. *In preparation*.

[48] J.B. Tenenbaum, V. de Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, **290**:2319–2323, 2000.

[49] R. Tibshirani. Principal curves revisited. *Statistics and Computing*, **2**:183–190, 1992.

[50] M. Trosset. Applications of multidimensional scaling to molecular conformation. *Computing Science and Statistics*, (29):148–152, 1998.

[51] L.J.P. van der Maaten, E.O. Postma, and H.J. van den Herik. Dimensionality reduction: a comparative review. 2007.

[52] Kilian Q. Weinberger and Lawrence K. Saul. Unsupervised learning of image manifolds by semidefinite programming. In *CVPR (2)*, pp. 988–995, 2004.

[53] Kilian Q. Weinberger, Fei Sha, and Lawrence K. Saul. Learning a kernel matrix for nonlinear dimensionality reduction. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, p. 106, New York, NY, USA, 2004. ACM.

[54] K.Q. Weinberger and L.K. Saul. An introduction to nonlinear dimensionality reduction by maximum variance unfolding. In *AAAI*, 2006.

[55] K.Q. Weinberger and L.K. Saul. Unsupervised learning of image manifolds by semidefinite programming. *International Journal of Computer Vision*, **70**(1):77–90, 2006.

[56] Li Yang. Building connected neighborhood graphs for isometric data embedding. In *KDD*, pp. 722–728, 2005.

[57] P. Zhand, Y. Huang, S. Shekhar, and V. Kumar. Correlation analysis of spatial time series datasets. In *Pacific Asia Conf. on Knowledge Discovery and Data Mining*, 2003.

[58] Hao Zhang, Oliver van Kaick, and Ramsay Dyer. Spectral mesh processing. *Computer Graphics Forum (to appear)*, 2008.

[59] A. Amadei, A.B.M. Linssen, and H.J.C. Berendsen. Essential dynamics of proteins. *Proteins: Structure, Function, and Genetics*, **17**(4):412–425, 1993.

[60] K.D. Ball, R.S. Berry, R.Kunz, F-Y. Li, A. Proykova, and D.J. Wales. From topographies to dynamics on multidimensional potential energy surfaces of atomic clusters. *Science*, 271(5251):963 – 966, 1996.

[61] O. Becker and M. Karplus. The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. *The Journal of Chemical Physics*, **106**(4):1495–1517, 1997.

[62] O.M. Becker. Principal coordinate maps of molecular potential energy surfaces. *J. of Comp. Chem.*, **19**(11):1255–1267, 1998.

[63] R. Stephen Berry, Nuran Elmaci, John P. Rose, and Benjamin Vekhter. Linking topography of its potential surface with the dynamics of folding of a proteinmodel. *Proceedings of the National Academy of Sciences*, **94**(18):9520–9524, 1997.

[64] Robert B. Best and Gerhard Hummer. Chemical Theory and Computation Special Feature: Reaction coordinates and rates from transition paths. *Proceedings of the National Academy of Sciences*, **102**(19):6732–6737, 2005.

[65] P.G. Bolhuis, D. Chandler, C. Dellago, and P.L. Geissler. Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annual review of physical chemistry*, **53**:291–318, 2002.

[66] P.G. Bolhuisdagger, C. Dellago, and D. Chandler. Reaction coordinates of biomolecular isomerization. *PNAS*, **97**(11):5877–5882, 2000.

[67] C.L. Brooks, J. Onuchic, and D.J. Wales. Statistical thermodynamics: taking a walk on a landscape. *Science*, **293**(5530):612 – 613, 2001.

[68] L. Chavez, J.N. Onuchic, and C. Clementi. Quantifying the roughness on the free energy landscape: Entropic bottlenecks and protein folding rates. *J. Am. Chem. Soc.*, **126**(27):8426–8432, 2004.

[69] Samuel S. Cho, Yaakov Levy, and Peter G. Wolynes. P versus Q: Structural reaction coordinates capture protein folding on smooth landscapes. *Proceedings of the National Academy of Sciences*, **103**(3):586–591, 2006.

[70] P. Das, M. Moll, H. Stamati, L. Kavraki, and C. Clementi. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *PNAS*, **103**(26):9885–9890, 2006.

[71] Payel Das, Corey J. Wilson, Giovanni Fossati, Pernilla Wittung-Stafshede, Kathleen S. Matthews, and Cecilia Clementi. Characterization of the folding landscape of monomeric lactose repressor: Quantitative comparison of theory and experiment. *Proceedings of the National Academy of Sciences*, **102**(41):14569–14574, 2005.

[72] R. Du, V. Pande, A.Y. Grosberg, T. Tanaka, and E.I. Shakhnovich. On the transition coordinate for protein folding. *J. Chem. Phys.*, **108**(1):334–350, 1998.

[73] R.L. Dunbrack. Rotamer libraries in the 21st century. *Curr. Opin. Struct. Biol.*, **12**(4):431–440, 2002.

[74] H.A. Scheraga et al. A united-residue force field for off-lattice protein-structure simulations. i. functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. *J. of Computational Chemistry*, **18**(7):849–873, 1997.

[75] A. Fersht. *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*. 1999.

[76] A.T. Fomenko and T.L. Kunii. *Topological Modeling for visualization*. Springer, 1997.

[77] D. Frenkel and B. Smit. *Understanding molecular simulation*. Academic Press, 2002.

[78] A.E. Garcia. Large-amplitude nonlinear motions in proteins. *Physical Review Letters*, **68**(17):2696–2699, 1992.

[79] D. Gfeller, P. De Los Rios, A. Caflisch, and F. Rao. Complex network analysis of free-energy landscapes. *Proceedings of the National Academy of Sciences*, **104**(6):1817–1822, 2007.

[80] Nobuhiro Go and Hiroshi Taketomi. Respective Roles of Short- and Long-Range Interactions in Protein Folding. *Proceedings of the National Academy of Sciences*, **75**(2):559–563, 1978.

[81] Isaac A. Hubner, Eric J. Deeds, and Eugene I. Shakhnovich. Understanding ensemble protein folding at atomic detail. *Proceedings of the National Academy of Sciences*, **103**(47):17747–17752, 2006.

[82] G. Hummer. From transition paths to transition states and rate coefficients. *J. Chemical Physics*, **120**(2), 2004.

[83] T. Ichiye and M. Karplus. Collective motions in proteins: A covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins: Structure, Function, and Genetics*, **11**(3):205–217, 1991.

[84] C.L Brooks III, M. Gruebele, J. Onuchic, and P. Wolynes. Chemical physics of protein folding. *Proceedings of the National Academy of Sciences*, **95**(19):11037–11038, 1998.

[85] S.E. Jackson. How do small single-domain proteins fold? *Fold Des.*, **3**(4):R81–91, 1998.

[86] J. Janin, S. Wodak, M. Levitt, and B. Maigret. Conformations of amino acid side chains in proteins. *J. Mol. Biol.*, **125**:357–386, 1978.

[87] T. Komatsuzaki, K. Hoshino, Y. Matsunaga, G.J. Rylance, R.L. Johnston, and D. Wales. How many dimensions are required to approximate the potential energy landscape of a model protein? *J. Chem. Phys.*, **122**, February 2005.

[88] R.E. Kunz and R.S. Berry. Statistical interpretation of topographies and dynamics of multidimensional potentials. *J. Chem. Phys.*, **103**:1904–1912, August 1995.

[89] O.F. Lange and H Grubmller. Generalized correlation for biomolecular dynamics. *Proteins*, **62**:1053–1061, 2006.

[90] C. Levinthal. Are there pathways for protein folding? *Journal de Chimie Physique et de Physico-Chimie Biologique*, **65**:44–45, 1968.

[91] John W. Milnor. *Morse Theory*. Princeton University Press, Princeton, NJ, 1963.

[92] E. Paci, M. Vendruscolo, and M. Karplus. Native and non-native interactions along protein folding and unfolding pathways. *Proteins*, **47**(3):379–392, 2002.

[93] J. Palis and W. de Melo. *Geometric Theory of Dynamical Systems*. Springer, 1982.

[94] M. Pettini. *Geometry and Topology in Hamiltonian Dynamics and Statistical Mechanics*. Springer, 2007.

[95] E. Plaku, H. Stamati, C. Clementi, and L.E. Kavraki. Fast and reliable analysis of molecular motion using proximity relations and dimensionality reduction. *Proteins: Structure, Function, and Bioinformatics*, **67**(4):897–907, 2007.

[96] G. Rylance, R. Johnston, Y. Matsunaga, C-B Li A. Baba, and T. Komatsuzaki. Topographical complexity of multidimensional energy landscapes. *PNAS*, **103**(49):18551–18555, 2006.

[97] M. Tirion. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.*, **77**:1905–1908, 1996.

[98] Monique M. Tirion. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.*, **77**(9):1905–1908, Aug 1996.

[99] D.J. Wales. *Energy Landscapes*. Cambridge University Press, 2003.

[100] L. Yang, G. Song, and R. Jernigan. Comparison of experimental and computed protein anisotropic temperature factors. In *IEEE Bioinformactics and biomedecine workshop*, 2007.

RATIONAL PARAMETRIZATIONS, INTERSECTION THEORY, AND NEWTON POLYTOPES*

CARLOS D'ANDREA† AND MARTÍN SOMBRA‡

Abstract. The study of the Newton polytope of a parametric hypersurface is currently receiving a lot of attention both because of its computational interest and its connections with Tropical Geometry, Singularity Theory, Intersection Theory and Combinatorics. We introduce the problem and survey the recent progress on it, with emphasis in the case of curves.

Key words. Parametric curve, implicit equation, Newton polytope, tropical geometry, Intersection Theory.

AMS(MOS) 2000 subject classifications. Primary 14Q05; Secondary 12Y05, 52B20, 14C17.

1. Introduction. Parametric curves and surfaces play a central role in Computer Aided Geometric Design (CAGD) because they provide shapes which are easy to plot. Indeed, a rational parametrization allows to produce many points in the variety using only the elementary operations $(\pm, \times, \div)$ of the base field.

For instance, consider the *folium of Descartes* (Figure 1). This plane curve can be defined either by the equation $x^3 + y^2 - 3xy = 0$ or as the image of the rational map

$$\mathbb{C} \dashrightarrow \mathbb{C} \quad , \quad t \mapsto \left(\frac{3t}{1+t^3}, \frac{3t^2}{1+t^3} \right). \tag{1}$$

The parametric representation is certainly more suitable for plotting the curve. If instead we plot it using only its implicit equation, the result is bound to be poor, specially around the singular point $(0, 0)$ (Figure 2).

This is because in order to produce many points in the folium in this way, we have to solve as many cubic equations. This is certainly more expensive than evaluating the parametrization and moreover, the resulting points are typically not rational but live in different cubic extensions of $\mathbb{Q}$.

On the other hand, if we are to decide whether a given point lies in the folium or not, it is better to use the implicit equation. For instance, it is straightforward to conclude that $(-2, 1)$ does not belong to the folium

*D'Andrea is partially supported by the Programa Ramón y Cajal of the Ministerio de Ciencia e Innovación (Spain) and by the research project MTM2007–67493. Sombra is partially supported by the research project MTM2006-14234.

†Universitat de Barcelona, Departament d'Àlgebra i Geometria, Gran Via 585, 08007 Barcelona, Spain (cdandrea@ub.edu, http://carlos.dandrea.name).

‡Université de Bordeaux 1, Institut de Mathématiques de Bordeaux, 351 cours de la Libération, 33405 Talence Cedex, France (martin.sombra@math.u-bordeaux1.fr, http://www.math.u-bordeaux1.fr/∼sombra).

I.Z. Emiris et al. (eds.), *Nonlinear Computational Geometry*, The IMA Volumes in Mathematics and its Applications 151, DOI 10.1007/978-1-4419-0999-2_2,

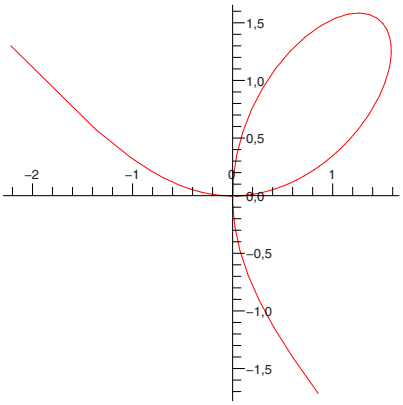


FIG. 1. *The folium of Descartes.*

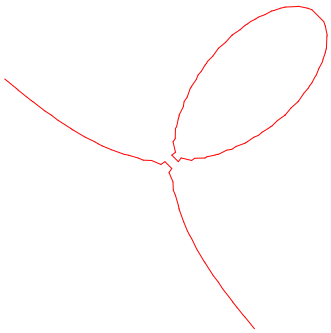


FIG. 2. *The folium of Descartes according to the Maple command* `implicitplot`.

by evaluating the equation: $(-2)^3 + 1^3 - 3(-2) = -1 \neq 0$. If we were to find that out from the parametrization (1), we would have to determine if the system of equations

$$-2 = \frac{3t}{1+t^3}, \qquad 1 = \frac{3t^2}{1+t^3}$$

admits a solution for $t \in \mathbb{C}$ or not, which is a harder task.

Depending on which kind of operation one needs to perform on a certain parametric variety, it may be convenient to dispose of the parametric representation or of the implicit one. Efficiently performing the passage from one representation to the other is one of the central problems of Computational Algebraic Geometry. In the present text we will mostly concentrate in one these directions: the *implicitization problem*, consisting in computing equations for an algebraic variety given in parametric form.

In precise terms, the implicitization problem is: let $\rho_1, \ldots, \rho_n \in \mathbb{C}(t_1, \ldots, t_{n-1})$ be a family of rational functions and consider the map

$$\boldsymbol{\rho} : \mathbb{C}^{n-1} \dashrightarrow \mathbb{C}^n \quad , \quad \boldsymbol{t} = (t_1, \ldots, t_{n-1}) \mapsto (\rho_1(\boldsymbol{t}), \ldots, \rho_n(\boldsymbol{t})). \tag{2}$$

Suppose that the Zariski closure $\overline{\mathrm{Im}(\rho)}$ of the image of this map is a hypersurface or equivalently, that the Jacobian matrix $(\frac{\partial \rho_i}{\partial t_j}(\boldsymbol{t}))_{i,j}$ has maximal rank $n-1$ for generic $\boldsymbol{t} \in \mathbb{C}^{n-1}$. The ideal of this parametric (or *unirational*) hypersurface is generated by a single irreducible polynomial and the problem consists in computing this "implicit equation".

This problem is equivalent to the elimination of the parameter variables from some system of equations. For instance, to compute the implicit equation of the folium from the parametrization (1), one should eliminate the variable t from the system of equations

$$(1+t^3)\,x - 3t = 0 \quad , \quad (1+t^3)\,y - 3t^2 = 0, \tag{3}$$

that is, we have to find the irreducible polynomial in $\mathbb{C}[x, y]$ vanishing at the points (x, y, t) satisfying (3) for some $t \in \mathbb{C}$.

The same procedure works in general. For a parametrization like in (2), write $\rho_i(\boldsymbol{t}) = \frac{p_i(\boldsymbol{t})}{q_i(\boldsymbol{t})}$ for some coprime polynomials p_i, q_i for $1 \leq i \leq n$. The implicit equation of the hypersurface $\overline{\mathrm{Im}(\boldsymbol{\rho})}$ can then be obtained by eliminating the variables $t_1, \ldots, t_{n-1}$ from the system of equations

$$q_1(\boldsymbol{t})x_1 - p_1(\boldsymbol{t}) = 0, \ldots, q_n(\boldsymbol{t})x_n - p_n(\boldsymbol{t}) = 0.$$

This elimination task can be effectively done either with Gröbner bases or with resultants [3] but in practice, this can be too expensive. For instance, for $a \in \mathbb{N}$ consider the parametrization

$$\boldsymbol{\rho} : \mathbb{C} \to \mathbb{C}^2 \quad , \quad t \mapsto \left(\frac{t(t-1)^a}{(t+1)^{a+1}}, \frac{(t+1)^a}{t(t-1)^{a-1}} \right). \tag{4}$$

It is not hard to check by hand that the implicit equation of the image of this map is

$$2 - x^{a-1}y^a - x^a y^{a+1} = 0.$$

However, all current implementations of Gröbner bases and resultant algorithms fail to solve the problem for moderately large values of a, because of the increasing number of intermediate computations involved.

2. The Newton polytope of the implicit equation. Instead of trying to compute the implicit equation of a parametric hypersurface, we will focus in the problem of determining its Newton polytope. We will work with *Laurent polynomials*, that is expressions of the form $x_2^{-1} + x_1^{-2}x_2$ where the exponents can be any integer numbers.

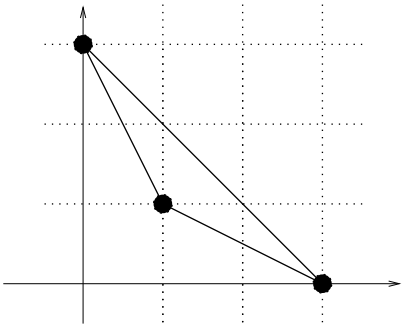


FIG. 3. *The Newton polygon of the folium of Descartes.*

DEFINITION 2.1. *The Newton polytope* $\mathrm{N}(F) \subset \mathbb{R}^n$ *of a Laurent polynomial* $F \in \mathbb{C}[x_1^{\pm 1}, \dots, x_n^{\pm 1}]$ *is the convex hull of the exponents in its monomial expansion.*

This notion readily extends to hypersurfaces: we define the *Newton polytope* $\mathrm{N}(Z)$ of a hypersurface $Z \subset \mathbb{C}^n$ as the Newton polytope of its implicit equation; this polytope is well defined because the equation is unique up to a scalar factor. For the case $n = 2$ we will apply the more usual terminology of "polygon" instead of polytope. For instance, the Newton polygon of the folium $x_1^3 + x_2^3 - 3x_1x_2 = 0$ is the convex hull $\mathrm{Conv}\big((1,1),(3,0),(0,3)\big)$ (Figure 3).

The Newton polytope tells us which are the possible exponents occuring in a given Laurent polynomial: if the polytope is small, then the polynomial is *sparse*, in the sense that it has few monomials. It is an important refinement of the notion of degree: if we denote by $S := \mathrm{Conv}(\mathbf{0}, e_1, \dots, e_n)$ the standard simplex of $\mathbb{R}^n$, the degree of a polynomial $F \in \mathbb{C}[x_1, \dots, x_n]$ is the least integer d such that $\mathrm{N}(F) \subset dS$. Note that the Newton polytope of a polynomial (and *a fortiori* that of a hypersurface) is always contained in the octant $(\mathbb{R}_{\geq 0})^n$.

From now on, we will focus in the following problem: determine the Newton polytope of a hypersurface given by a rational map $\boldsymbol{\rho} : \mathbb{C}^{n-1} \dashrightarrow \mathbb{C}^n$. This problem is currently receiving a lot of attention because of its connections with Tropical Geometry, Singularity Theory, Intersection Theory and Combinatorics. The Newton polytope does not characterize the hypersurface but retains a lot of relevant information and as a consequence of the research done during the last years, we now know that in plenty of cases its computation is much simpler than that of the full implicit equation.

A preliminary version of this question was first posed by B. Sturmfels and J.-T. Yu. In the context of the sparse elimination theory, their question can be resumed in: "can I predict the Newton polytope of the implicit equation from the Newton polytopes of the input parametrization?" In precise terms:

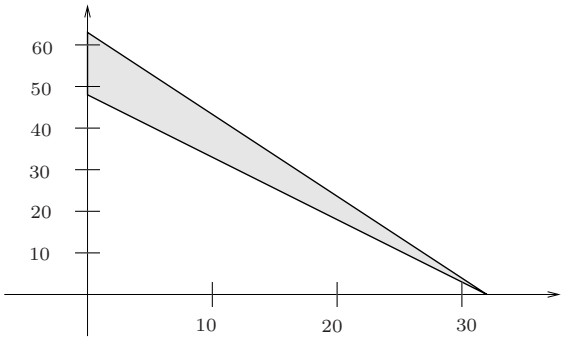


FIG. 4. *The Newton polygon of the implicit equation of (5).*

PROBLEM 2.2. *Let $P_1, \ldots, P_n \subset \mathbb{R}^{n-1}$ be lattice polytopes with non-empty interior and consider the family of n Laurent polynomials in $n-1$ variables*

$$\rho_i = \sum_{a \in P_i \cap \mathbb{Z}^{n-1}} \lambda_{i,a}\, t_1^{a_1} \cdots t_{n-1}^{a_{n-1}} \in \mathbb{C}\Big[t_1^{\pm 1}, \ldots, t_{n-1}^{\pm 1}\Big]$$

for $1 \leq i \leq n$ and $\lambda_{i,a} \in \mathbb{C}$ generic. Determine the Newton polytope of the image of the parametrization $\boldsymbol{t} \mapsto (\rho_1(\boldsymbol{t}), \ldots, \rho_n(\boldsymbol{t}))$.

A *lattice polytope* in $\mathbb{R}^{n-1}$ is a polytope whose vertices lie in $\mathbb{Z}^{n-1}$. The hypothesis that the P_i's have non empty interior ensures that the image of the parametrization is a hypersurface for a generic choice of the coefficients $\lambda_{i,a}$ (that is, in some non empty open set of the space of parameters). The Newton polytope of this hypersurface does not depend on this generic choice although the equation itself does.

As an example, let us consider the parametrization proposed by A. Dickenstein and R. Fröberg [8]:

$$\boldsymbol{\rho} : \mathbb{C} \to \mathbb{C}^2 \quad , \quad t \mapsto \big(t^{48} - t^{56} - t^{60} - t^{62} - t^{63}, t^{32}\big). \tag{5}$$

The Newton polytopes of the defining polynomials are relatively small: the real interval $[48, 63]$ and the singleton $\{32\}$. The exponents are rather large, but in any case the implicit equation can be computed *via* the Sylvester resultant. It's Newton polygon is the triangle with vertices $(32, 0)$, $(0, 48)$, $(0, 63)$.

This example was studied by I. Emiris and I. Kotsireas, who succeeded in determining the polygon by analysing the behavior of the resultant under specialization, thus showing that it is sometimes possible to access to the Newton polytope without computing the implicit equation [8]; see also [7] for further applications of this method.

The recent irruption of Tropical Geometry in the mathematical panorama has boosted the interest in the problem. The tropical variety

associated to an affine hypersurface is a polyhedral object, equivalent to its Newton polytope in the sense that one can be recovered from the other and viceversa. In this direction, Sturmfels, J. Tevelev and J. Yu succeeded in determining the tropical variety (and thus the Newton polytope) of a hypersurface parametrized by generic Laurent polynomials [15, 16] and implemented the resulting algorithm [17, 18].

From another point of view, A. Esterov and A. Khovanskiĭ have shown that the Newton polytope of the implicit equation of a generic parametrization can be identified with the *mixed fiber polytope* in the sense of P. McMullen, hence providing a different characterization of this object [9].

2.1. The Newton polygon of a parametric curve. If one wants to determine the Newton polytope in *all* cases and not just in the generic ones, it is clear that finer invariants of the parametrization must be taken into account.

In this section we will focus in the case of parametric plane curves, which has been recently solved in the papers [4, 5, 15, 20]. In this case, the Newton polygon is determined by the multiplicities of the parametrization. Let $\rho : \mathbb{C} \dashrightarrow \mathbb{C}^2$ be a map given by rational function $f, g \in \mathbb{C}(t) \setminus \mathbb{C}$. For a point v in the projective line $\mathbb{P}^1$, the *multiplicity of* ρ *in* v is

$$\operatorname{ord}_v(\rho(t)) := \big(\operatorname{ord}_v(f(t)), \operatorname{ord}_v(g(t))\big) \in \mathbb{Z}^2,$$

where $\operatorname{ord}_v(f)$ denotes the order of vanishing of f at v. Recall that the order of vanishing at $v = \infty$ of a rational function $\frac{p}{q} \in \mathbb{C}(t)$ $(p, q \in \mathbb{C}[t])$ equals $\deg(q) - \deg(p)$.

The basic properties of these multiplicities are:

- $\operatorname{ord}_v(\rho) = (0, 0)$ except for a finite number of $v \in \mathbb{P}^1$ and
- $\sum_{v \in \mathbb{P}^1} \operatorname{ord}_v(\rho) = (0, 0)$.

We next define an auxiliary operation which produces a convex lattice polygon from a balanced family of vectors of the plane. Let $B \subset \mathbb{Z}^2$ be a family of vectors which are zero except for a finite number of them and such that $\sum_{b \in B} b = (0, 0)$. We denote by $\mathcal{P}(B) \subset (\mathbb{R}_{\geq 0})^2$ the (unique) convex polygon obtained by: 1) rotating $-90°$ the non-zero vectors of B, 2) concatenating them following their directions counterclockwise and 3) translating the resulting polygon to the first quadrant $(\mathbb{R}_{\geq 0})^2$ in such a way that it "touches" the coordinate axes (Figure 5). The zero-sum condition warrants that the polygonal line closes at the end of the concatenation step.

The *tracing index* (or *degree*) $\operatorname{ind}(\rho) \geq 1$ is the number of times the parametrization ρ runs over the curve when t runs over $\mathbb{C}$. When $\operatorname{ind}(\rho) = 1$, we say that the parametrization is *birational*.

The solution to the problem of the computation of the Newton polygon of a parametric plane curve can be found in the papers of Dickenstein, E.-M. Feichtner, Sturmfels and Tevelev [5, 15, 20] and also in ours [4].

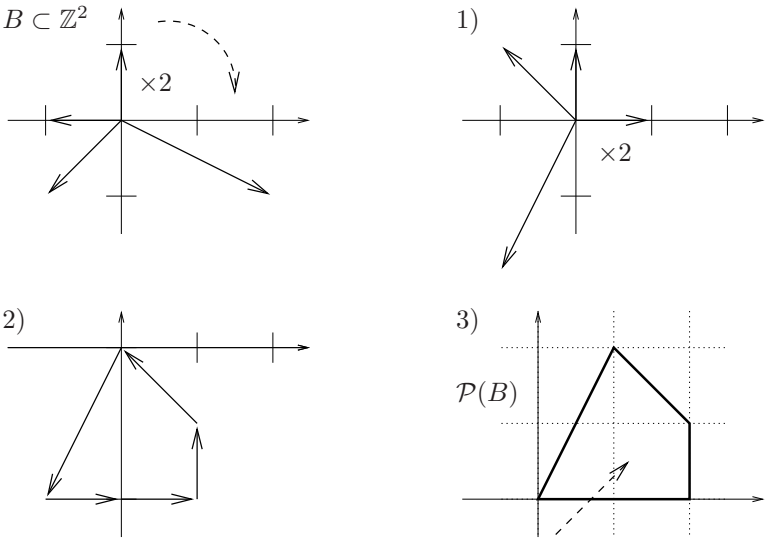


FIG. 5. *The operation* $\mathcal{P}(B)$.

THEOREM 2.1. *Let* $\rho : \mathbb{C} \dashrightarrow \mathbb{C}^2$ *be a rational map and set* $C := \overline{\mathrm{Im}(\rho)}$, *then*

$$\mathrm{N}(C) = \frac{1}{\mathrm{ind}(\rho)} \mathcal{P}\big((\mathrm{ord}_v(\rho))_{v\in\mathbb{P}^1}\big). \tag{6}$$

EXAMPLE 1. *Consider the parametrization*

$$\rho : t \mapsto \Big(\frac{1}{t(t-1)}, \frac{t^2-5t+2}{t}\Big).$$

Its multiplicities are

$$\mathrm{ord}_0(\rho) = (-1,-1) \quad , \quad \mathrm{ord}_1(\rho) = (-1,0) \quad , \quad \mathrm{ord}_\infty(\rho) = (2,-1)$$

and $\mathrm{ord}_{v_i}(\rho) = (0,1)$ *for each of the two zeros* v_1, v_2 *of the equation* $t^2 - 5t + 2 = 0$, *while* $\mathrm{ord}_v(\rho) = (0,0)$ *for* $v \neq 0, \pm 1, \infty, v_1, v_2$. *Figure 5 illustrates the family* B *and the associated polygon* $\mathcal{P}(B)$.

Theorem 2.1 tells us that this polygon is $\mathrm{ind}(\rho)$ *times the actual Newton polygon of the curve. It is easy to check that the constructed polygon is non contractible, in the sense that it is not a non trivial integer multiple of another lattice polygon. We conclude that the parametrization is birational (that is,* $\mathrm{ind}(\rho) = 1$*) and that* $\mathrm{N}(C) = \mathcal{P}(B)$. *These results can be*

contrasted with the implicit equation of the curve: $1 - 16x - 4x^2 - 9xy - 2x^2y - xy^2 = 0$.

Similarly, Theorem 2.1 allows to determine the Newton polygon of the folium of Descartes and of the Dickenstein-Fröberg example. Again, the additional data $\mathrm{ind}(\rho) = 1$ is a consequence of the non contractibility of the resulting polygon.

2.2. Tropical geometry and Intersection Theory. We will sketch here the two different methods for proving Theorem 2.1: Tropical Geometry and Intersection Theory.

Tropical Geometry can be regarded as the geometry of the min-plus algebra $(\mathbb{R}, \oplus, \odot)$, where the operations are defined as

$$x \oplus y = \min(x, y) \quad , \quad x \odot y = x + y.$$

To simplify the exposition, we will only deal with polynomials in $\mathbb{C}[x, y]$, although the theory extends naturally to multivariate polynomials with coefficients in a valuated field.

The *tropicalization* of a polynomial $F = \sum_{j=0}^{N} \lambda_j x^{a_j} y^{b_j} \in \mathbb{C}[x, y]$ is the concave piecewise linear function

$$t_F : \mathbb{R}^2 \to \mathbb{R} \quad , \quad x \mapsto \bigoplus_{j=0}^{N} x^{\odot a_j} y^{\odot b_j} = \min_j \langle (a_j, b_j), (x, y) \rangle. \tag{7}$$

Here, $\bigoplus$ stands for the tropical sum and $\langle \cdot, \cdot \rangle$ is the standard inner product of $\mathbb{R}^2$. The *tropical variety* $\mathcal{T}_F \subset \mathbb{R}^2$ is defined as the set of points in $\mathbb{R}^2$ where this function is not smooth. It can be deduced from (7) that $\mathcal{T}_F$ consists exactly in the union of the outer normal directions to the edges of $\mathrm{N}(F)$. To each of these directions δ we can assign a *multiplicity* $m_\delta \geq 1$, which coincides with the lattice length of the edge of $\mathrm{N}(F)$ normal to the given direction. We recall that the *lattice length* $\ell(S)$ of a lattice segment S is the number of points in $\mathbb{Z}^2 \cap S$ minus 1.

This setting allows to interpret the Newton polygon as a certain degeneration of the curve and to study it with tools of Tropical Geometry. The proof given in [5,12] of Theorem 2.1 is based in the so-called "Kapranov Theorem" [6] and the Bieri-Groves Theorem. Moreover, their method allows them to treat higher-dimensional hypersurfaces parametrized by products of linear forms [5, 15].

As an illustration, Figure 6 shows the tropical variety associated to the curve in the example 1:

As we can see, the tropical variety plus the corresponding multiplicities are in correspondence with the vectors in Figure 5 and Theorem 2.1 can be easily reformulated in tropical terms.

On the other hand, in our paper [4] we propose a method which reduces the determination of the Newton polygon to the computation of the number of solutions of some polynomial systems of equations.

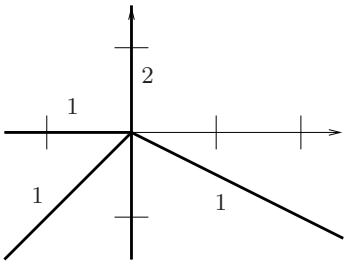


FIG. 6. *The tropical curve associated with* $C : 1 - 16x - 4x^2 - 9xy - 2x^2y - xy^2 = 0$.

The *support function* of a polygon $Q \subset \mathbb{R}^2$ is defined as

$$h_Q : \mathbb{R}^2 \to \mathbb{R} \quad , \quad x \mapsto \max\{\langle u, x\rangle : u \in Q\}.$$

It is a convex piecewise affine function which completely characterizes Q. Let $\rho : \mathbb{C} \dashrightarrow \mathbb{C}^2$ be a rational parametrization and set $C := \overline{\mathrm{Im}(\rho)}$. Then, for $\sigma \in (\mathbb{N} \setminus \{0\})^2$ it can be shown that

$$h_{\mathrm{N}(C)}(\sigma) = \frac{1}{\mathrm{ind}(\rho)} \#\big\{(t,x,y) \in \mathbb{C}^3 : x^{\sigma_1} = f(t), y^{\sigma_2} = g(t), \\ \ell_0 + \ell_1 x + \ell_2 y = 0\big\} \tag{8}$$

for generic $\ell_0, \ell_1, \ell_2 \in \mathbb{C}$. The proof of Theorem 2.1 reduces then to the determination of this number of solutions, which can be obtained *via* the refinement of the Bernštein-Kušnirenko-Khovanskiĭ (BKK) Theorem recently obtained by P. Philippon and the second author [13].

Identity (8) holds also in higher dimensions. However, there is no analogue for $n \geq 3$ of the estimation in [13] and so for the moment, this method cannot be extended to higher dimension.

3. Some applications and consequences. Besides of its theoretical interest, the Newton polytope is useful for computational purposes. Its knowledge allows to speed-up computations and gives interesting information about the solutions of polynomial systems of equations. Here we point out two applications.

3.1. Computing the implicit equation with numerical interpolation. The Newton polytope tells us which exponents might occur in the implicit equation and thus allows us to compute it *via* a suitable interpolation algorithm. Suppose we are given a parametrization $\boldsymbol{\rho} = (f, g) : \mathbb{C} \dashrightarrow \mathbb{C}^2$ and that we want to compute the implicit equation $E(x, y) \in \mathbb{C}[x, y]$ of its image curve. A possible strategy is to apply Theorem 2.1 to obtain its Newton polygon Q and use this information to recover E. We have

$$E(x,y) = \sum_{j=0}^{N} \lambda_j x^{a_j} y^{b_j}$$

where the (a_j, b_j)'s are the integer points in Q and the $\lambda_j \in \mathbb{C}$ are unknown. To determine these coefficients, we can evaluate $\boldsymbol{\rho}$ in $N+1$ points $\tau_0, \ldots, \tau_N \in \mathbb{C}$ where $\rho(\tau_i)$ is defined. We then obtain a homogeneous system of linear equations in the λ_j's, of size $(N+1) \times (N+1)$:

$$E(\boldsymbol{\rho}(\tau_k)) = \sum_{j=0}^{N} \lambda_j f(\tau_k)^{a_j} g(\tau_k)^{b_j} = 0 \quad \text{for } 0 \leq k \leq N.$$

If the interpolation points τ_k are generic enough, the solution space of this system is of dimension 1 and the polynomial $E(x,y)$ can be computed as some (any) generator of this space. This approach is most useful when Q has few points integer points, which for instance is the case of the parametrization (4), where the number of integer points is 3 for any $a \in \mathbb{N}$.

3.2. Intersecting parametric curves. In the practice of CAGD, it is important to be able to determine where two modelled shapes cut each other. Typically, this amounts to compute the intersection of two curves or surfaces given in parametric form. This task can be done by computing the implicit equation of one of the two varieties but as explained, this can be too expensive. If we only have access to the Newton polytope, we will not be able to compute this intersection but we still can say something about the number of intersection points of two parametric curves or about the degree of the intersection curve of two parametric surfaces.

For two plane curves $C, D \subset \mathbb{C}^2$, the BKK Theorem says that the number of their intersection points in $(\mathbb{C}^\times)^2$ is bounded above by the *mixed volume*

$$\mathrm{Area}(\mathrm{N}(C) + \mathrm{N}(D)) - \mathrm{Area}(\mathrm{N}(C)) - \mathrm{Area}(\mathrm{N}(D))$$

with equality in the generic case. Here, the "+" denotes the Minkowski (that is, pointwise) sum of polygons in the plane. For instance, let $C, D \subset (\mathbb{C}^\times)^2$ be the curves respectively parametrized by

$$t \mapsto \Big(\frac{(t+1)^2}{2\,t\,(1-t)}, \frac{4\,t\,(t-1)^3}{(t+1)^5}\Big) \quad , \quad t \mapsto \Big(t, \frac{10}{t^3}\Big).$$

In Figure 7 we see the corresponding polygons and their Minkowski sum. The mixed volume is the area of the shaded zone, which is equal to 2.

Hence C and D have at most two points in common, which turn out to be $(1.33, 4.22)$ and $(-4.17, -0.14)$ (Figure 8).

3.3. Generic parametrizations. With Theorem 2.1 at our disposal, we can easily answer Problem 2.2 for the case $n = 2$:

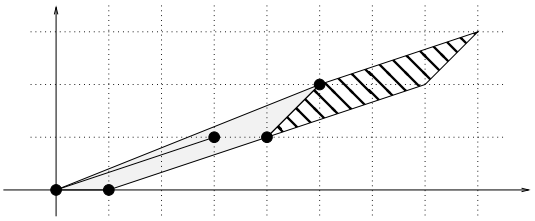


FIG. 7. *The mixed volume of two polygons.*

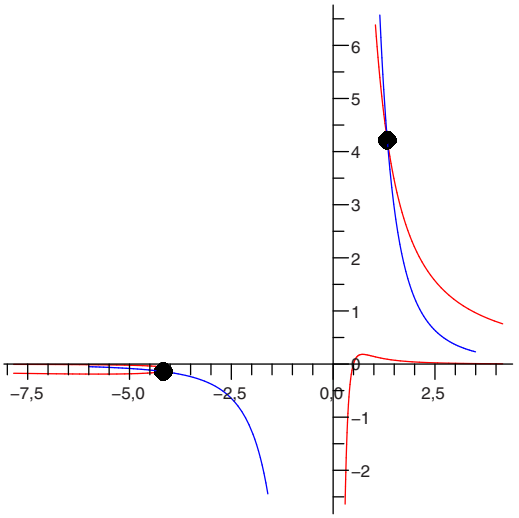


FIG. 8. *The intersection of two parametric curves.*

COROLLARY 3.1. *For $D \geq d$, $E \geq e$ let*

$$p(t) = \alpha_d t^d + \cdots + \alpha_D t^D \quad , \quad q(t) = \beta_e t^e + \cdots + \beta_E t^E \quad \in \mathbb{C}[t^{\pm 1}] \quad (1)$$

such that $\alpha_d, \alpha_D, \beta_e, \beta_E \neq 0$ and $\gcd(t^{-d}p(t), t^{-e}q(t)) = 1$. Set $\rho = (p, q)$ and $C := \overline{\mathrm{Im}(\rho)}$, then

$$\mathrm{N}(C) = \frac{1}{\mathrm{ind}(\rho)} \mathcal{P}\big((D-d, 0), (0, E-e), (-D, -E), (d, e)\big).$$

In particular, parametrizations by generic Laurent polynomials produce equations whose Newton polygon is typically a quadrilateral (Figure 9).

The proof of this corollary is simple: we have $\mathrm{ord}_0(\rho) = (d, e)$ and $\mathrm{ord}_\infty(\rho) = (-D, -E)$. Let $v_1, \ldots, v_r \neq 0$ be the different roots of $t^{-d}p(t)$ and $m_i \geq 1$ the multiplicity of v_i in p, then

$$\mathrm{ord}_{v_i}(\rho) = (m_i, 0) \quad \text{for } 1 \leq i \leq r$$

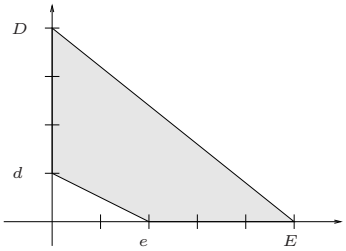


FIG. 9. *The Newton polygon of a generic Laurent polynomial parametrization.*

as we assume that p and q do not share roots in the torus. Similarly, let $w_1, \dots, w_s \neq 0$ be the roots of $t^{-e}q(t)$ and $n_j \geq 1$ be their respective multiplicities. For the same reasons as above

$$\operatorname{ord}_{w_j}(\rho) = (0, n_j).$$

Theorem 2.1 then shows that $\operatorname{ind}(\rho)\,\mathrm{N}(C)$ is obtained by rotating $-90°$ and concatenating the vectors (d, e), $(-D, -E)$, $(m_i, 0)$ and $(0, n_j)$, for $1 \leq i \leq r$ and $1 \leq j \leq s$. But the $(m_i, 0)$'s are all pointing in the same direction and so they concatenate together into the vector $\sum_i (m_i, 0) = (D - d, 0)$. Similarly, the $(0, n_j)$'s concatenate together into $\sum_j (0, n_j) = (0, E - e)$, which concludes the proof.

Moreover, it can be shown that for a parametrization like (1), the Newton polygon of the implicit equation equals

$$\frac{1}{\operatorname{ind}(\rho)} \mathcal{P}\big((D - d, 0), (0, E - e), (-D, -E), (d, e)\big)$$

if and only if $\alpha_d, \alpha_D, \beta_e, \beta_E \neq 0$ and $\gcd(t^{-d}p(t), t^{-e}q(t)) = 1$. If besides the vectors $(D-d, 0), (0, E-e), (d, e)$ are not collinear, then ρ is birational.

Note that the polygon does not depend on the actual values of the roots of p and q, it only depends on the hypothesis that they are disjoint and that we know the sum of their multiplicities. This is a general principle: for computing the Newton polygon of a parametrization $\rho = (f, g)$, we do not need full access to the zeros and poles of f and g. It suffices with partial factorizations of the form

$$f(t) = \alpha \prod_{p \in P} p(t)^{d_p} \quad , \quad g(t) = \beta \prod_{p \in P} p(t)^{e_p}$$

where $\mathcal{P} \subset \mathbb{C}[t]$ is a finite set of relatively prime polynomials, $d_p, e_p \in \mathbb{N}$ and $\alpha, \beta \in \mathbb{C}^\times$. Such factorizations can be obtained with gcd's operations only, which is certainly easier than extracting roots and poles.

Back to the world of generic parametrizations, the second case to tackle is when we have two rational functions with the same denominator. It turns out that the resulting Newton polygon has at most five edges (Figure 10).

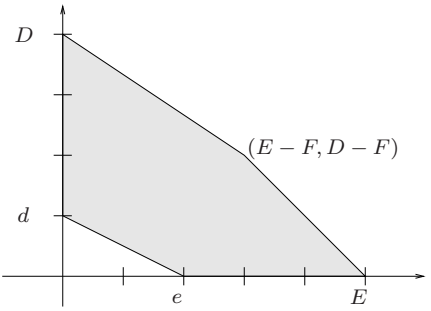


FIG. 10. *The Newton polygon of a parametrization by generic rational functions with the same denominator.*

COROLLARY 3.2. *Given $D \geq d$, $E \geq e$ and $F \geq 0$, let*

$$p(t) = \alpha_d t^d + \cdots + \alpha_D t^D, \quad q(t) = \beta_e t^e + \cdots + \beta_E t^E, \quad r(t) = \gamma_0 + \cdots + \gamma_F t^F.$$

Set $\rho = \left(\frac{p}{r}, \frac{q}{r}\right) \in \mathbb{C}(t)^2$ and $C := \overline{\mathrm{Im}(\rho)}$, then

$$\mathrm{N}(C) = \frac{1}{\mathrm{ind}(\rho)} \mathcal{P}\big((D-d,0),(0,E-e),(F-D,F-E),(d,e),(-F,-F)\big)$$

if and only if $\alpha_d, \alpha_D, \beta_e, \beta_E, \gamma_0, \gamma_F \neq 0$ and $t^{-d}p(t), t^{-e}q(t), r(t)$ are pairwise coprime.

Finally, we consider the case when the parametrization is given by two generic rational functions with different denominators. The resulting polygon has at most six edges (Figure 11).

COROLLARY 3.3. *Given $D \geq d$, $E \geq e$, $F, G \geq 0$, let*

$$p(t) = \alpha_d t^d + \cdots + \alpha_D t^D \quad , \quad q(t) = \beta_e t^e + \cdots + \beta_E t^E \quad \in \mathbb{C}[t^{\pm 1}]$$

and

$$r(t) = \gamma_0 + \cdots + \gamma_F t^F \quad , \quad s(t) = \delta_0 + \cdots + \delta_G t^G \quad \in \mathbb{C}[t].$$

Set $\rho = \left(\frac{p}{r}, \frac{q}{s}\right)$ and $C := \overline{\mathrm{Im}(\rho)}$, then

$$\mathrm{N}(C) = \frac{1}{\mathrm{ind}(\rho)} \mathcal{P}\big((D-d,0),(0,E-e),(F-D,G-E),(d,e),(-F,0),(0,-G)\big)$$

if and only if $\alpha_d, \alpha_D, \beta_e, \beta_E, \gamma_0, \gamma_F, \delta_0, \delta_G \neq 0$ and $t^{-d}p(t), t^{-e}q(t), r(t), s(t)$ are pairwise coprime.

4. The general case *vs* the generic case. Now suppose we start from the other endpoint, that is suppose that we are given the equation $E(x, y)$ of a parametric curve. What does its Newton polytope tell us about the (unknown) parametrization?

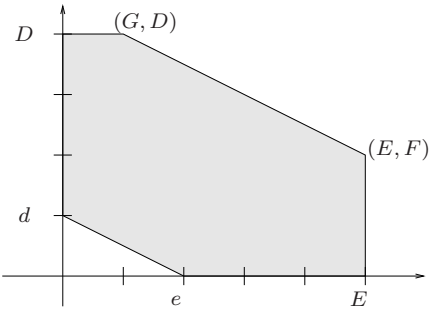


FIG. 11. *The Newton polygon of a generic parametrization by rational functions with different denominators.*

A first natural question is whether $\mathrm{N}(E(x,y))$ can be *any* lattice polygon. As we have seen, the polygons produced by generic parametrizations are very special: they have at most six edges and some of them are in prefixed directions.

Before answering this question, let us fix a lattice polygon $Q \subset (\mathbb{R}_{\geq 0})^2$ with non empty interior and touching the coordinate axes. We will identify $\mathbb{C}^{\#(Q\cap\mathbb{Z}^2)}$ with the $\mathbb{C}$-vector space of polynomials whose Newton polygon is contained in Q. Consider the set

$$M_Q^\circ := \big\{F \in \mathbb{C}[x,y] :\ \mathrm{N}(F) = Q,\ F \text{ defines a parametric curve in } \mathbb{C}^2\big\} \subset \mathbb{C}^{\#(Q\cap\mathbb{Z}^2)}$$

and let M_Q denote its Zariski closure. Recall that ∂Q denotes the *border* of Q.

THEOREM 4.1 ([4]). *M_Q is a parametric variety of dimension $\#(\partial Q \cap \mathbb{Z}^2)$.*

In particular, $\dim(M_Q) \geq 3$ as Q must have at least three edges. It turns out that any lattice polygon with non empty interior and supported in the coordinate axes is the Newton polygon of a parametric curve.

A further consequence of this result is that the codimension of M_Q equals the number of lattice points in the interior of Q. This is interesting for the inverse problem: given a polynomial $E(x,y) \in \mathbb{C}[x,y]$, decide whether it defines a parametric curve or not and if it is the case, compute a parametrization.

If the Newton polygon of the equation has a lot of points in its interior, then the probability that E defines a parametric curve is low. If nevertheless this is the case, the corresponding parametrization will be defined by $\#(\partial Q\cap\mathbb{Z}^2)$ degrees of freedom, and hence the efficiency of the computation of such a parametrization should be correlated with the number of lattice points in ∂Q and not with the number of lattice points in the whole of Q.

Some pointers to the literature:

- Parametric curves in general: [1, 14, 21]
- Numerical interpolation methods: [2, 10, 19]
- Newton polytopes and specialized resultants: [7, 8]
- Newton polytopes and Tropical Geometry: [5, 12, 15, 16, 17]
- Newton polytopes and mixed fiber polytopes: [9, 11, 17]
- Newton polytopes and Intersection Theory: [4, 13]

Acknowledgments. We are grateful to the Institute for Mathematics and its Applications (IMA) and specially to the organizers of the Workshop on Non-Linear Computational Geometry held there during the Spring 2007, where this project was started.

REFERENCES

[1] S.S. Abhyankar, *Algebraic geometry for scientists and engineers*, Amer. Math. Soc., 1990.

[2] R.M. Corless, M.W. Giesbrecht, I. Kotsireas, and S.M. Watt, Numerical implicitization of parametric hypersurfaces with linear algebra, *Lecture Notes in Comput. Sci.* **1930** (2001) 174–183.

[3] D. Cox, J. Little, and D. O'Shea, *Using algebraic geometry. Second edition*, Graduate Texts in Math. **185**, Springer, 2005.

[4] C. D'Andrea and M. Sombra, The Newton polygon of a rational plane curve, e-print `arXiv:0710.1103`, 26pp..

[5] A. Dickenstein, E.-M. Feichtner, and B. Sturmfels, Tropical discriminants, *J. Amer. Math. Soc.* **20** (2007) 1111–1133.

[6] M. Einsiedler, M. Kapranov, and D. Lind, Non-Archimedean amoebas and tropical varieties, *J. Reine Angew. Math.* **601** (2006) 139–157.

[7] I. Emiris, C. Konaxis, and L. Palios, Computing the Newton polygon of the implicit equation, e-print `arXiv:0811.0103`, 21 pp..

[8] I. Emiris and I. Kotsireas, Implicitization exploiting sparseness, *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.* **67** (2005) 281–298.

[9] A. Esterov and A. Khovanskiĭ, Elimination theory and Newton polytopes, *Funct. Anal. Other Math.* **2** (2008) 45–71.

[10] A. Marco and J.J. Martínez, Implicitization of rational surfaces by means of polynomial interpolation, *Comput. Aided Geom. Design* **19** (2002) 327–344.

[11] P. McMullen, Mixed fibre polytopes, *Discrete Comput. Geom.* **32** (2004) 521–532.

[12] G. Mikhalkin, Tropical geometry and its applications, *International Congress of Mathematicians* Vol. II, 827–852, Eur. Math. Soc., 2006.

[13] P. Philippon and M. Sombra, A refinement of the Bernštein-Kušnirenko estimate, *Adv. Math.* **218** (2008) 1370–1418.

[14] J.R. Sendra, F. Winkler, and S. Pérez-Díaz, *Rational Algebraic Curves. A Computer Algebra Approach*, Springer, 2008.

[15] B. Sturmfels and J. Tevelev, Elimination theory for tropical varieties, *Math. Res. Lett.* **15** (2008) 543–562.

[16] B. Sturmfels, J. Tevelev, and J. Yu, The Newton polytope of the implicit equation, *Moscow Mathematical Journal* **7** (2007) 327–346.

[17] B. Sturmfels and J. Yu, *Tropical implicitization and mixed fiber polytopes*, in M. Stillman, N. Takayama, and J. Verschelde (eds.), *Software for Algebraic Geometry*, I.M.A. Volumes in Mathematics and its Applications **148** (2008) 111–132.

[18] B. Sturmfels and J. Yu, TrIm: a software for tropical implicitization, `http://www-math.mit.edu/~jyu/TrIm/`.

[19] Y. Sun and J. Yu, Implicitization of parametric curves via Lagrange interpolation, *Computing* **77** (2006) 379–386.

[20] J. Tevelev, Compactifications of subvarieties of tori, *Amer. J. Math.* **129** (2007) 1087–1104.

[21] R.J. Walker, *Algebraic Curves*, Princeton Univ. Press, 1950.

SOME DISCRETE PROPERTIES OF THE SPACE OF LINE TRANSVERSALS TO DISJOINT BALLS

XAVIER GOAOC*

Abstract. Attempts to generalize Helly's theorem to sets of lines intersecting convex sets led to a series of results relating the geometry of a family of sets in $\mathbb{R}^d$ to the structure of the space of lines intersecting all of its members. We review recent progress in the special case of disjoint Euclidean balls in $\mathbb{R}^d$, more precisely the inter-related notions of *cone of directions*, *geometric permutations* and *Helly-type theorems*, and discuss some algorithmic applications.

Key words. Geometric transversal, Helly's theorem, line, sphere, geometric permutation, cone of directions.

1. Introduction. Lines intersecting or tangent to prescribed geometric objects are central to various problems in computational geometry and application areas; typical examples include visibility [26, 64] or shortest path [61] computation and robust statistics [14, 67]. To design efficient algorithms for these problems, one first has to understand the geometry of the underlying sets of lines. A natural embedding of the space of lines in $\mathbb{P}^3(\mathbb{R})$ is as a quadric in $\mathbb{P}^5(\mathbb{R})$, the Klein (or Plücker) quadric; in some sense this is optimal[1], so line geometry is, at least in dimension 3, inherently nonlinear.

Let $\mathcal{C}$ be a collection of subsets of $\mathbb{R}^d$, or *objects* for short. Denote by $\mathcal{T}_k(\mathcal{C})$ the set of k-transversals to $\mathcal{C}$, that is of k-dimensional affine subspaces that intersect every member of $\mathcal{C}$. Helly's theorem [42] asserts that if $\mathcal{C}$ consists of convex sets then $\mathcal{T}_0(\mathcal{C})$ is nonempty if and only if $\mathcal{T}_0(F)$ is nonempty for any subset $F \subset \mathcal{C}$ of size at most $d+1$. Whether Helly's theorem generalizes to other values of k is a natural question which was, to my knowledge, first investigated in the 1930's by Vincensini [76]. The answer turns out to be negative in general but positive when the geometry of the objects is adequately constrained. The study of how the geometry of the objects in $\mathcal{C}$ determines the structure of $\mathcal{T}_k(\mathcal{C})$, and subsequent developments of similar flavor, is now designated as *geometric transversal theory* [34].

Helly's theorem was recently generalized to *line transversals* ($k = 1$) to disjoint (Euclidean) balls in $\mathbb{R}^d$, answering in the positive a conjecture of Danzer [23] who settled the 2-dimensional case. This generalization builds on a series of results concerning two notions: *cone of directions* and *geometric permutations*. This survey gives a comprehensive overview of these investigations by presenting, in a unified language, the results of

*Loria - INRIA Nancy Grand-Est. (goaoc@loria.fr).

[1]Indeed (i) there does not exist any homeomorphism between the lines in $\mathbb{R}^3$ and an open subset of $\mathbb{P}^4(\mathbb{R})$, and (ii) any algebraic homeomorphism between lines in $\mathbb{R}^3$ and points in $\mathbb{P}^5(\mathbb{R})$ has degree at least 2 [65, Remarks 2.1.4 and 2.1.6, p. 143].

I.Z. Emiris et al. (eds.), *Nonlinear Computational Geometry*, The IMA Volumes in Mathematics and its Applications 151, DOI 10.1007/978-1-4419-0999-2_3,

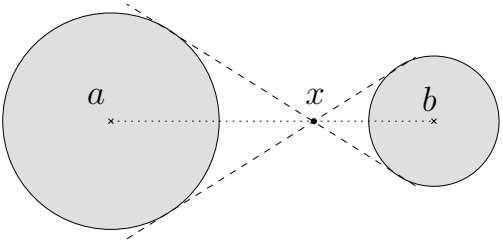


FIG. 1. *The internal center of similitude x of two balls, represented in a 2-plane through their centers.*

several papers [3, 7, 16, 17, 19–21, 45, 48, 53, 73, 81], new extensions of these results and some of their algorithmic consequences. Although some results generalize to other settings, the discussion will focus on the case of line transversals to disjoint balls.

1.1. Notations and terminology. We denote by $\mathbb{R}^d$ the real d-dimensional affine space or, equivalently, the Euclidean d-dimensional space; the Euclidean metric is the only one we consider over $\mathbb{R}^d$. We denote by $\mathbb{P}^d$ the real d-dimensional projective space and by $\mathbb{S}^{d-1}$ the space of directions in $\mathbb{R}^d$, which we identify with the unit sphere. Recall that a *great circle* of $\mathbb{S}^d$ is a section of $\mathbb{S}^d$ by some 2-flat through its center. We write A^o the interior of a set A and use arrows to denote vectors; in particular, we write $\overrightarrow{\ell}$ a direction vector of an oriented line ℓ. We use $\langle \overrightarrow{u}, \overrightarrow{v} \rangle$ and $\angle(\overrightarrow{u}, \overrightarrow{v})$ to denote, respectively, the dot product of and the angle between vectors $\overrightarrow{u}$ and $\overrightarrow{v}$.

A *ball* is closed unless otherwise specified: the ball of center c and radius r is the set of points x such that $|c - x| \leq r$. In particular, *disjoint* balls are not allowed to be tangent; for the sake of simplicity, we say that several balls are *disjoint* if they are pairwise disjoint. A *unit* ball is a ball with radius 1; since transversal properties are unchanged under scaling, results obtained for unit balls usually extend to *congruent* balls, i.e. sets of balls with equal radii. The *radius disparity* of a set of balls is the ratio of the largest radius to the smallest. The *internal center of similitude* of two disjoint balls in $\mathbb{R}^d$ with respective centers a, b and radii r_a, r_b is the point $\frac{r_b a + r_a b}{r_a + r_b}$ (see Figure 1); this point is sometimes referred to as the *geometric center* [81] or the *center of gravity* [47] of the two balls.

We use the terms *collection* or *family* for an unordered set, and *sequence* for an ordered set. We denote by $|X|$ the cardinality of a set X. Given a sequence $\mathcal{C}$, we denote by $\prec_{\mathcal{C}}$ the corresponding ordering on its elements. A *subsequence* of a sequence is a subset of its members, ordered as in the sequence. A *k-transversal* to a collection $\mathcal{C}$ is an affine subspace of dimension k that intersects every member of $\mathcal{C}$; for the sake of simplicity, we say *transversal* for 1-transversal, that is line transversal, and speak

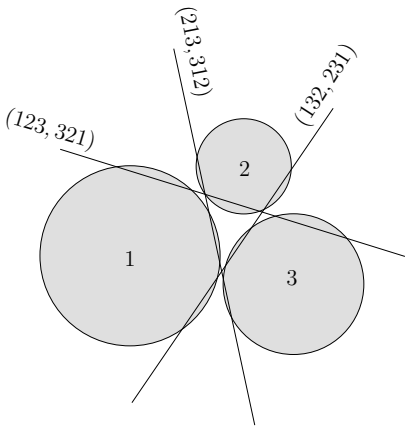


FIG. 2. *Three disks with three geometric permutations.*

of the common intersection of a family for the common intersection of its members. Depending on the context, a transversal may be oriented or not. An oriented transversal to a collection $\mathcal{C}$ of convex sets induces an ordering on $\mathcal{C}$, and a partial ordering if some of its members intersect. A *geometric permutation* of $\mathcal{C}$ is a pair of orderings, one reverse of the other, on $\mathcal{C}$ induced by the two orientations of some transversal (see Figure 2). A transversal to a sequence $\mathcal{C}$ is *order-respecting* if it meets the members of $\mathcal{C}$ in the order $\prec_{\mathcal{C}}$. The directions of all order-respecting transversals to $\mathcal{C}$ make up a subset of $\mathbb{S}^{d-1}$ called the *cone of directions* of $\mathcal{C}$ and denoted $\mathcal{K}(\mathcal{C})$.

The *projection along* a direction $\overrightarrow{u}$ is the orthogonal projection on some hyperplane with normal $\overrightarrow{u}$; since we consider properties invariant under translation of the hyperplane, all such hyperplanes are equivalent.

A *pinning configuration* is a ordered pair $(\mathcal{C}, \ell)$ where $\mathcal{C}$ is a collection of objects having ℓ as an isolated transversal, in the sense that ℓ is an isolated point of $\mathcal{T}_1(\mathcal{C})$. Equivalently we say that $\mathcal{C}$ *pins* the line ℓ. A pinning configuration $(\mathcal{C}, \ell)$ is *minimal* if no proper subset of $\mathcal{C}$ pins ℓ.

1.2. Content and organization. We start by recalling, in Section 2, some variants of Helly's theorem used in the rest of the paper. We then discuss, in Section 3, the convexity of the connected components of the projection of $\mathcal{T}_1(\mathcal{C})$ in the space of directions, the so-called cones of directions, when $\mathcal{C}$ consists of disjoint balls; we present the two existing approaches to proving this result [3, 16, 19, 40, 45] and discuss some of its immediate consequences. In Section 4, we review the bounds obtained on the number of geometric permutations of disjoint balls [7, 20, 21, 48, 53, 73, 81]. Section 5 then discusses Helly-type theorems for line transversals to disjoint balls. Specifically, we present the known bounds on the constants k for which the following statements hold:

(a) a sequence of disjoint balls has a transversal if every subsequence of size at most k has an order-respecting transversal,
(b) a collection of disjoint balls (with adequate constraints on the radii) has a transversal if every subset of size at most k has a transversal,
(c) if a transversal to n disjoint balls is isolated then it is an isolated transversal to a subset of at most k of the balls.

The smallest such constants are referred to as, respectively, the *Hadwiger*, *Helly*[2] and *pinning* numbers. Section 6 reviews the connection between Helly-type theorems and LP-type problems [4] and the impact of the bounds on the Helly number on the computational complexity of finding a transversal to a family of disjoint balls. We conclude this paper by commenting some open problems in Section 7.

1.3. Related surveys. For an overview of geometric transversal theory we refer to the surveys of Wenger [79] and Goodman et al. [35]. A detailed account on early generalizations of Helly's theorem can be found in the article by Danzer et al. [24] and more recent developments are presented in the survey of Eckhoff [27]. The overlap of the present survey has with the, related, ones of Sottile and Theobald [74] and Holmsen [44] is limited, so they can be read in conjunction. For a discussion of the computational aspects of lines in space we refer to the survey of Pellegrini [64] and the notes of the Alcala lecture by Pach and Sharir [63, Chapter 7].

2. Helly's theorem. Helly's theorem [42] of 1923 [3] forms, together with Radon's and Caratheodory's theorems, the basis of convex geometry [57].

THEOREM 2.1. *A finite family of $n \geq d+1$ convex sets in $\mathbb{R}^d$ has a point in common if and only if every $d+1$ members have a point in common.*

One way to restate Helly's theorem is that the emptiness of the intersection of any finite number of convex sets in $\mathbb{R}^d$ can be decided by looking only at subsets of size at most $d+1$. Other results of similar flavor are called *Helly-type theorems*; the typical formulation is that a collection $\mathcal{C}$ has property $\mathcal{P}$ if and only if every subset of $\mathcal{C}$ of size at most k has property $\mathcal{P}$ (k being independent of $|\mathcal{C}|$). The smallest integer k for which a given theorem holds is called the associated *Helly number*. In this section, we review some of these results used when dealing with transversals.

2.1. Spherical Helly theorem. There are several generalizations of Helly's theorem on the sphere $\mathbb{S}^d$ involving various notions of convexity [66]. Recall that a set $A \subset \mathbb{S}^d$ is *strongly convex* if it does not contain any pair of antipodal points and if it contains for any two points in A the smallest great circle arc that connects them. A strongly convex set $A \subset \mathbb{S}^d$ is *strictly*

[2] That is, the Helly number for sets of line transversals in the classical sense.

[3] It is sometimes dated from 1913, the year when Helly communicated it to Radon [24].

strongly convex if any great circle intersects its boundary in at most two points.

THEOREM 2.2. *A finite family of $n \geq d+2$ strongly convex sets in $\mathbb{S}^d$ has a point in common if and only if every $d+2$ members have a point in common.*

Proof. Consider a family $\mathcal{C}$ of strongly convex sets on a sphere $\mathbb{S}^d$ embedded in $\mathbb{R}^{d+1}$ with center O. Replacing each set $X \in \mathcal{C}$ by $X' = CH(X \cup \{O\}) \setminus \{O\}$, where $CH(\cdot)$ denotes the convex hull operator, we get a family $\mathcal{C}'$ of convex sets in $\mathbb{R}^{d+1}$ that has a common intersection if and only if $\mathcal{C}$ has a common intersection. The statement follows. □

With additional constraints on the sets one may reduce the Helly number to $d+1$. One simple example is if all sets in the family are contained in some open hemisphere of $\mathbb{S}^d$, as one can map that hemisphere to $\mathbb{R}^d$ while preserving the convexity structure. Another situation of interest is when the diameter[4] of the sets is bounded [66, Theorem 3]:

THEOREM 2.3. *A finite family of $n \geq d+2$ convex sets in $\mathbb{S}^d$, each of diameter less than $\frac{2\pi}{3}$, has nonempty intersection if and only if every $d+1$ members have nonempty intersection.*

2.2. Topological Helly theorem. Helly's theorem still holds if convexity is replaced by some weaker topological condition. Recall that a homology cell is a nonempty set with trivial homology. In particular, we use the following variant [25] of Helly's topological theorem [43]:

THEOREM 2.4. *Let $\mathcal{C}$ be a finite family of open subsets of $\mathbb{R}^d$ such that the intersection of any r elements of $\mathcal{C}$ is a homology cell for $r \leq d$. Then all sets in $\mathcal{C}$ have a point in common if and only if every $d+1$ members do.*

Recall that a set is *contractible* if it is homotopic to a point. The above theorem remains true if "homology cell" is replaced by "contractible set" since contractible subsets of $\mathbb{R}^d$ are homology cells (homology being invariant under homotopy); we will, in fact, only need this simpler variant.

2.3. Helly's theorem for unions of sets. The previous theorems do not apply to families of disconnected sets, as they are neither convex nor homologically trivial. Helly's theorem does, however, generalize to collections such that any members intersect in a bounded number of convex sets. The following theorem was conjectured by Grünbaum and Motzkin [39] and proven by Amenta [5]:

THEOREM 2.5. *Let $\mathcal{C}$ be a collection of sets in $\mathbb{R}^d$ such that the intersection of any nonempty finite sub-family of $\mathcal{C}$ is the disjoint union of at most k closed convex sets. Then all sets in $\mathcal{C}$ have a point in common if and only if every $k(d+1)$ members do.*

The same argument as in the proof of Theorem 2.2 yields:

[4] The diameter of $X \subset \mathbb{S}^d$ is the maximal opening angle of any great circle arc contained in X.

COROLLARY 2.1. *Let $\mathcal{C}$ be a collection of sets in $\mathbb{S}^d$ such that the intersection of any nonempty finite sub-family of $\mathcal{C}$ is the disjoint union of at most k closed convex sets. Then all sets in $\mathcal{C}$ have a point in common if and only if every $k(d+2)$ members do.*

Similar generalizations were obtained for the topological versions of Helly's theorem [2, 58]. We use the following corollary of a theorem of Matoušek [58, Theorem 2]:

THEOREM 2.6. *For any $d \geq 2$, $k \geq 1$ there exists a number $h(d,k)$ such that the following holds. Let $\mathcal{C}$ be a collection of sets in $\mathbb{R}^d$ such that the intersection of any nonempty finite sub-family of $\mathcal{C}$ has at most k path-connected components, each of them contractible. Then $\mathcal{C}$ has a point in common if and only if every $h(k,d)$ members have a point in common.*

2.4. Convexity structure on the Grassmannian. Transversals to convex sets provide an elegant way to define a "convexity" structure on the Grassmaniann [31, 33]: convex sets of k-flats are simply defined as the sets of k-transversals to convex objects. When $k = 1$ and the objects are restricted to axis-aligned boxes, the resulting structure is known as *frame convexity*. In fact, frame convexity, when restricted to ascending lines[5], is isomorphic to the ordinary notion of convexity on some convex subset of $\mathbb{R}^{2d-2}$; through this isomorphism, Helly's theorem essentially corresponds to Santaló's theorem [68], one of the earliest Helly-type theorems for line transversals [32]. As the examples of Santaló [68] and Danzer [23] show, Helly's theorem does not extend to the more general convexity structure of Goodman and Pollack [33].

3. Cone of directions. One of the specificities of the set of transversals to disjoint balls [6] is the following convexity property [16, Theorem 1]:

THEOREM 3.1. *The cone of directions of any sequence of disjoint balls in $\mathbb{R}^d$ is a strictly, strongly convex subset of $\mathbb{S}^{d-1}$.*

The use of the convexity of the cone of directions for proving Helly-type theorems for line transversals can be traced back to Vincensini [76]. In dimension 3 or more, Theorem 3.1 was first asserted[7] for the case of *thinly distributed* families of balls [40], i.e. families where the distance between the centers of any two balls is at least twice the sum of their radii.

3.1. Reduction. We first explain why Theorem 3.1 follows from the case of 3 balls in 3 dimensions.

LEMMA 3.1. *If $\mathcal{C}$ is a sequence of disjoint balls in $\mathbb{R}^d$ then $\mathcal{K}(\mathcal{C})$ is convex if $\mathcal{K}(X \cap T)$ is convex for any triple $X \subset \mathcal{C}$ and any 3-transversal T to X.*

[5] A line is ascending with respect to a coordinate frame if it can be oriented so that all coordinates are nondecreasing.

[6] This property implies that the Hadwiger number is bounded (see Lemma 5.1), which is not the case for disjoint translates of a convex set [46].

[7] Although Hadwiger's article does not contain a proof of this claim, its editor seems to have been provided with the details [3].

Proof. The statement follows from two facts: (i) $\mathcal{K}(\mathcal{C})$ is convex if $\mathcal{K}(\mathcal{C} \cap T)$ is convex for every 3-transversal T of $\mathcal{C}$ and (ii) $\mathcal{K}(\mathcal{C})$ is convex if $\mathcal{K}(X)$ is convex for any subsequence $X \subset \mathcal{C}$ of size d.

Let ℓ and ℓ' be two order-respecting transversals to a sequence $\mathcal{C}$ of disjoint convex sets in $\mathbb{R}^d$ and T some[8] 3-space that contains their span. Observe that T is a 3-transversal to $\mathcal{C}$ whose vector space contains any direction in the great circle spanned by $\overrightarrow{\ell}$ and $\overrightarrow{\ell'}$. Thus, if $\mathcal{K}(\mathcal{C} \cap T)$ is convex then $\mathcal{K}(\mathcal{C})$ contains the shorter arc of the great circle between $\overrightarrow{\ell}$ and $\overrightarrow{\ell'}$. As a consequence, we get that if $\mathcal{K}(\mathcal{C} \cap T)$ is convex for every 3-transversal to $\mathcal{C}$, then $\mathcal{K}(\mathcal{C})$ is convex. This proves claim (i).

Applying Helly's theorem to the projection of $\mathcal{C}$ along some direction $\overrightarrow{u}$, we find that $\mathcal{C}$ has a transversal with direction $\overrightarrow{u}$ if and only if every subset of size d has a transversal with direction $\overrightarrow{u}$. Since two parallel lines intersect disjoint convex sets in the same order,

$$\mathcal{K}(\mathcal{C}) = \bigcap_{X \subset \mathcal{C}, |X|=d} \mathcal{K}(X),$$

and claim (ii) follows. □

This reduction holds more generally for sequences of disjoint convex sets.

3.2. Analytic approach. The first published proof of Theorem 3.1 came for families of disjoint unit balls in $\mathbb{R}^3$ [45]. This case was not covered by Hadwiger's statement since two unit balls within distance $\delta \in (0, 2)$ are disjoint but not thinly distributed.

Let $\mathcal{C}$ be a sequence of disjoint unit balls in $\mathbb{R}^3$ and $u, v \in \mathcal{K}(\mathcal{C})$. Consider some coordinate axis, say z, normal to these two directions and let R denote some plane parallel to z. Let $Q^{uv}_{\mathcal{C}} \subset \mathbb{R} \times \mathbb{S}^1$ denote the set of (t, α) such that there is an oriented line in the plane $z = t$ that intersects the sequence $\mathcal{C}$ in the right order and makes an angle α with R. The projection of $Q^{uv}_{\mathcal{C}}$ on the second coordinate is exactly the intersection of $\mathcal{K}(\mathcal{C})$ with the great circle through u and v; the convexity of $Q^{uv}_{\mathcal{C}}$ for all pairs $u, v \in \mathcal{K}(\mathcal{C})$ thus implies that of $\mathcal{K}(\mathcal{C})$. Let A and B be two balls such that $A \prec B$ in $\mathcal{C}$; define similarly $Q^{uv}_{AB} \subset \mathbb{R} \times \mathbb{S}^1$ as the set of all (t, α) such that there is an oriented line in the plane $z = t$ that intersects A before B and makes an angle α with R. Helly's theorem implies that

$$Q^{uv}_{\mathcal{C}} = \bigcap_{A \prec B \text{ in } \mathcal{C}} Q^{uv}_{AB},$$

so it suffices to prove the convexity of Q^{uv}_{AB}. Using symmetries with respect to translation and rotation, the convexity of Q^{uv}_{AB} reduces to the convexity of the function

[8] If ℓ and ℓ' are skew T is unique.

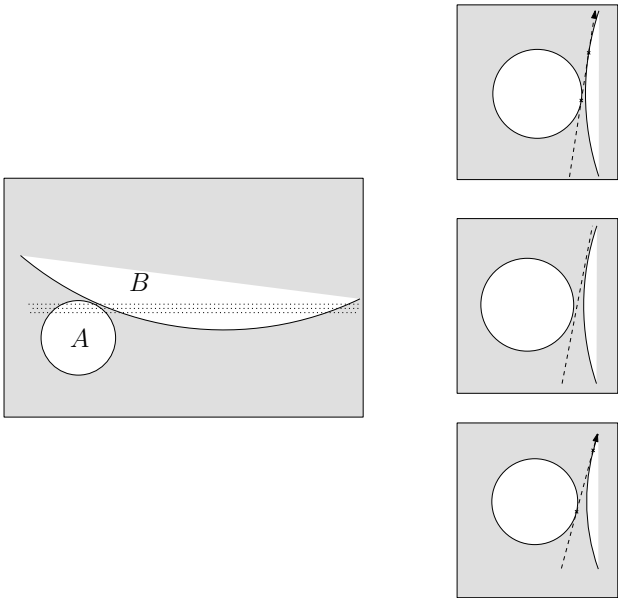


FIG. 3. *Two balls such that Q^{uv}_{AB} is not convex, in the (x, z) plane (left). Three slices at (x, y) planes (regularly spaced along the y axis) showing that the "middle" of two existing transversals is not a transversal (right).*

$$G : t \mapsto \arcsin\left(\frac{\sqrt{1-z^2}+\sqrt{1-(z-b)^2}}{a}\right),$$

on the interval $[\frac{b}{2}, 1]$, where a and b parameterize the respective positions of the centers. Elementary calculus suffices to conclude.

3.3. Extending the analytic approach. The convexity of Q^{uv}_{AB} is stronger than that of the cone of directions: it requires that if $\mathcal{C}$ has transversals with directions $\overrightarrow{u}$ and $\overrightarrow{v}$ in planes $z = z_u$ and $z = z_v$, then for any $t \in [0, 1]$ it has a transversal with direction $t\overrightarrow{u} + (1-t)\overrightarrow{v}$ *in the plane* $z = tz_u + (1-t)z_v$. This property does, in fact, not hold for disjoint balls with arbitrary radii; Figure 3 depicts an example of two 3-dimensional disjoint balls for which the nonconvexity of certain sets $Q^{uv}_{\mathcal{C}}$ can be ascertained [36]; the balls have centers $(0, 0, 0)$ and $(3.9, 0, 8.6)$ and radii 1 and 8.44. Of course, this nonconvexity may (and, in fact, does) disappear when $Q^{uv}_{\mathcal{C}}$ is projected on the second coordinate, so this does not disprove Theorem 3.1. It does, however, show that the previous approach requires a constraint stronger than the balls' disjointedness. That approach was, nevertheless, extended in two directions.

Ambrus et al. [3] used this technique to prove Theorem 3.1 for d-dimensional unit balls such that the distance between any two centers is at least $2\sqrt{2+\sqrt{2}}$. The key observation is that the distance between two

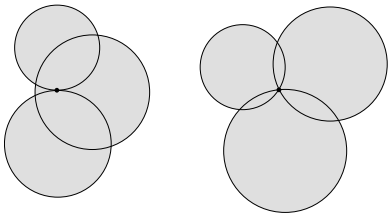


FIG. 4. *Intersection patterns for the projections of three balls along a direction on the boundary of their cone of directions: direction of inner special bitangent (left) and of tritangent(right).*

centers in the section of such a sequence by any of its 3-transversals is at least $2\sqrt{1+\sqrt{2}}$; this guarantees the convexity of the function G for these sections, and thus for the d-dimensional balls (by the argument used to prove Lemma 3.1).

Cheong et al. [19] proved Theorem 3.1 for sequences of balls where every pair is isometric to the section of two higher-dimensional disjoint unit balls by some d-transversal; such *pairwise-inflatable* pairs are characterized by the property that the squared distance between their centers is at least twice the sum of their squared radii[9]. The convexity of Q^{uv}_{AB} is first established for disjoint unit balls in $\mathbb{R}^4$ via extensive computations[10], then deduced for pairwise-inflatable balls in $\mathbb{R}^3$ and finally extended to pairwise-inflatable balls in $\mathbb{R}^d$.

3.4. The algebraic approach. The general case of Theorem 3.1 was proven by Borcea et al.[16] by showing that the algebraic arcs that make up the boundary of $\mathcal{K}(\mathcal{C})$ do not contain any inflexion point. By Lemma 3.1, it suffices to consider the case where $\mathcal{C}$ is a triple of disjoint balls in $\mathbb{R}^3$.

Boundary arcs. The directions that belong to the boundary of $\mathcal{K}(\mathcal{C})$ can be characterized in terms of projection patterns [19, Lemma 9 and 11]:

LEMMA 3.2. *A direction $\overrightarrow{u}$ belongs to the boundary of $\mathcal{K}(\mathcal{C})$ if and only if the projections of the balls of $\mathcal{C}$ along $\overrightarrow{u}$ intersect in a single point.*

An immediate consequence is that the directions in the interior of $\mathcal{K}(\mathcal{C})$ are exactly the directions of transversals to the open balls in $\mathcal{C}$:

$$\mathcal{K}(\mathcal{C}^o) = \mathcal{K}^o(\mathcal{C}).$$

The intersection of the projections of a triple $\mathcal{C}$ of balls in $\mathbb{R}^3$ along $\overrightarrow{u} \in \partial\mathcal{K}(\mathcal{C})$ belongs to the boundary of either two or three disks (see Figure 4). The boundary of $\mathcal{K}(\mathcal{C})$ thus decomposes into two types of arcs, directions

[9] This is, somewhat unexpectedly, simply Hadwiger's condition of being thinly-distributed where every distance is replaced by its square. Note that disjoint unit balls are pairwise-inflatable.

[10] Although computer algebra systems such as Maple [56] were instrumental in developing these computations, the resulting proof can still be checked manually.

of *inner special bitangents*, i.e. tangents to two balls through their inner center of similitude[11], and of tritangents. The directions of inner special bitangents make up a circle on $\mathbb{S}^2$, so the local convexity of arcs of the first type is trivial. The directions of tritangents make up an algebraic curve of degree 6 on $\mathbb{S}^2$, the *direction-sextic* of the triple of balls, that is not convex in general. It is thus important to identify the directions of tritangents that belong to the boundary of $\mathcal{K}(\mathcal{C})$ [16, Proposition 3]:

LEMMA 3.3. *The direction of a tritangent ℓ is on the boundary of the cone of directions of three balls (for the adequate ordering) if and only if ℓ intersects the triangle spanned by their centers.*

This generalizes to higher dimensions and follows from the property that the balls centered at the vertices of a simplex and going through a given point have no other common intersection if and only if that point belongs to the simplex.

Controlling the flexes. Proving Theorem 3.1 essentially amounts to showing that the boundary of the cone of directions of three disjoint balls does not contain inflexion points of the curve of directions of tritangents; these, also called *flexes*, are the intersections of the curve with its Hessian and the sources of nonconvexity in an algebraic curve. Given a projection pattern of a sequence $\mathcal{C}$ of three balls along some direction $\overrightarrow{u} \in \partial\mathcal{K}(\mathcal{C})$, the conditions that the Hessian of the direction-sextic of $\mathcal{C}$ vanishes in $\overrightarrow{u}$, and that the balls are disjoint exclude one another [16, Proposition 5]. This approach avoids the apparently difficult task of classifying the 72 flexes of the direction-sextic[12].

3.5. Strict convexity and tangents to spheres. The algebraic approach immediately yields that the cone of direction is *strictly convex*, in the sense that its boundary does not contain great circle arcs. This property is also related [19, Proposition 4] to collections of spheres with degenerate families of common tangents [74]. In $\mathbb{R}^3$, if the cone of directions of three balls contains a great circle arc then these balls have infinitely many common tangents that meet one and the same line at infinity [19, Lemma 10]. Such configurations require the balls to intersect [60], so the strict convexity follows for three, and hence n, disjoint balls in $\mathbb{R}^3$. The generalization to higher dimensions is based on the following lemma:

LEMMA 3.4. *For any sequence $\mathcal{C}$ of disjoint balls in $\mathbb{R}^d$ and any great circle $\Gamma \subset \mathbb{S}^{d-1}$ there exists a 3-transversal T to $\mathcal{C}$ such that $\mathcal{K}(\mathcal{C}) \cap \Gamma = \mathcal{K}(\mathcal{C} \cap T) \cap \Gamma$. Moreover, for any such 3-space we have $\partial\mathcal{K}(\mathcal{C}) \cap \Gamma \subset \partial\mathcal{K}(\mathcal{C} \cap T) \cap \Gamma$.*

Proof. Since $\mathcal{K}(\mathcal{C})$ is convex, its intersection with Γ is a (possibly empty) small great circle arc η. If η is reduced to a single point then Lemma 3.2 implies that $\mathcal{C}$ has a unique transversal with direction in Γ,

[11] These lines are exactly the tangents to two balls contained in a common tangent plane; they are sometimes referred to as *limiting bitangents*.

[12] This bound is tight if intersections are counted with multiplicities and over $\mathbb{P}^2(\mathbb{C})$.

and any 3-space T containing this transversal will do. Otherwise, let T be some[13] 3-space containing the two transversals to $\mathcal{C}$ with directions in $\partial\eta$. As $\mathcal{K}(\mathcal{C} \cap T)$ is convex, its intersection with Γ is a small great circle arc. Since this arc contains $\partial\eta$ it contains η, and the other inclusion is immediate.

Let T be a 3-transversal to $\mathcal{C}$ such that $\mathcal{K}(\mathcal{C}) \cap \Gamma = \mathcal{K}(\mathcal{C} \cap T) \cap \Gamma$. By Lemma 3.2, the projections of $\mathcal{C}$ along any direction $\overrightarrow{u} \in \partial\mathcal{K}(\mathcal{C}) \cap \Gamma$ intersect in a single point; the projections of $\mathcal{C} \cap T$ along $\overrightarrow{u}$ must then also intersect in a single point, and $\overrightarrow{u} \in \partial\mathcal{K}(\mathcal{C} \cap T)$. □

In particular, if the cone of directions of some sequence of disjoint balls $\mathcal{C}$ in $\mathbb{R}^d$ contains a great circle arc Γ on its boundary, then Γ also appears on the boundary of the section of $\mathcal{C}$ by some 3-transversal; the strict convexity thus extends from the 3-dimensional case to higher dimensions.

3.6. Immediate consequences. Let $\mathcal{C}$ be a finite collection of disjoint balls in $\mathbb{R}^d$. The following are simple consequences of Theorem 3.1.

3.6.1. Topology of order-respecting transversals. Obviously, two transversals to $\mathcal{C}$ that realize distinct geometric permutations belong to different connected components of $\mathcal{T}_1(\mathcal{C})$. Theorem 3.1 implies that the converse is true [19, Lemma 14]:

THEOREM 3.2. *The set of transversals to a finite number of disjoint balls in $\mathbb{R}^d$ in a given order is contractible.*

Proof. Let $\mathcal{C}$ be a finite sequence of disjoint balls and L its set of order-respecting transversals. A transversal ℓ to $\mathcal{C}$ is *barycentric* if it goes through the center of mass of the intersection of the projections of the balls in $\mathcal{C}$ along $\overrightarrow{\ell}$. For any direction v in $\mathcal{K}(\mathcal{C})$ there is a unique barycentric transversal to $\mathcal{C}$, which we denote $b_{\mathcal{C}}(v)$. Let L^* denote the set of order-respecting barycentric transversals to $\mathcal{C}$. The projection of a ball changes continuously with the direction of projection, so $b_{\mathcal{C}}$ is continuous. Since the direction of a line changes continuously with the line, $b_{\mathcal{C}}^{-1}$ is also continuous and $b_{\mathcal{C}}$ defines a homeomorphism between L^* and $\mathcal{K}(\mathcal{C})$. By Theorem 3.1, $\mathcal{K}(\mathcal{C})$ is convex and hence contractible. It follows that L^* is also contractible. The map

$$\begin{cases} L \times [0,1] \to L \\ (\ell, t) \mapsto \ell + t(b_{\mathcal{C}}(v_\ell) - \ell) \end{cases}$$

is continuous and shows that L^* is a deformation retract of L. Since L^* is contractible, so is L. □

3.6.2. Isotopy and geometric permutations. Two transversals to $\mathcal{C}$ are said to be *isotopic* if one can be moved continuously into the other while remaining a transversal during the motion, i.e. if they belong to the same path-connected component of $\mathcal{T}_1(\mathcal{C})$. Theorem 3.2 implies that the

[13] If the two transversals are skew, T is unique.

number of geometric permutations of $\mathcal{C}$ is equal to the number of connected components of $\mathcal{T}_1(\mathcal{C})$:

COROLLARY 3.1. *Two transversals to a finite family of disjoint balls in $\mathbb{R}^d$ are isotopic if and only if they induce the same geometric permutation.*

Koltun and Sharir [55, Theorem 5.4] showed that the number of isotopy classes of transversals to n disjoint balls is $O(n^{3+\epsilon})$ for $d = 3$ and $O(n^{2d-2})$ for $d \geq 4$; their proofs recast the set of transversals as a sandwich region in an arrangement of hyperplanes and builds on a series of results on the structure of such arrangements. With Corollary 3.1, Theorem 4.1 immediately improves these bounds.

3.6.3. Pinning configurations. Since $\mathcal{T}_1(\mathcal{C})$ can be recast as an union of cells in an arrangement of algebraic surfaces of bounded degree [55], it has a bounded number of connected components. Thus, a point in $\mathcal{T}_1(\mathcal{C})$ is isolated if and only if it is a connected component of $\mathcal{T}_1(\mathcal{C})$. Minimal pinning configurations can then be characterized as follows:

COROLLARY 3.2. *Let ℓ be an order-respecting transversal to a finite sequence $\mathcal{C}$ of disjoint balls in $\mathbb{R}^d$. $\mathcal{C}$ pins ℓ if and only if no other transversal to $\mathcal{C}$ realizes the same geometric permutation as ℓ, or equivalently:*

$$\mathcal{K}(\mathcal{C}) = \{\overrightarrow{\ell}\} \quad \Leftrightarrow \quad \mathcal{K}^o(\mathcal{C}) = \emptyset \quad \Leftrightarrow \quad \mathcal{K}(\mathcal{C}^o) = \emptyset.$$

Proof. Since $\mathcal{C}$ pins ℓ if and only if $\{\ell\}$ is a connected component of $\mathcal{T}_1(\mathcal{C})$, the first equivalence follows from Theorem 3.2. If $\mathcal{K}(\mathcal{C}) = \{\overrightarrow{\ell}\}$ then Lemma 3.2 ensures that no other line realizes the same geometric permutation as ℓ, and the second equivalence follows. The remaining equivalences are straightforward. □

Since a transversal is isolated if and only if no other transversal realizes the same geometric permutation, and two lines are always contained in some common 3-space, we have:

COROLLARY 3.3. *A finite collection $\mathcal{C}$ of disjoint balls in $\mathbb{R}^d$ pins a line ℓ if and only if for every 3-space T that contains ℓ, $\mathcal{C} \cap T$ pins ℓ in T.*

4. Geometric permutations. The first investigation of geometric permutations is, to the best of our knowledge, due to Katchalski, Lewis and Liu [50]. Since then, the maximum number of geometric permutations was studied for a variety of different shapes: convex sets [6, 10, 28, 52, 70, 78], boxes [80], fat convex sets [54], translates of a convex set [8, 9, 11, 51], balls [73], congruent balls [20, 21, 48, 53, 73], balls with bounded radius disparity [48, 81]. For disjoint balls, the bounds can be summarized as follows:

THEOREM 4.1. *The maximum number of geometric permutations of a family of n disjoint balls in $\mathbb{R}^d$ is $\Theta\left(n^{d-1}\right)$ if the balls have arbitrary radii, $O\left(\gamma^{\log \gamma}\right)$ if the balls have radius disparity at most γ, at most 3 if the balls have equal radii and at most 2 if, in addition, $n \geq 9$ or $n \geq 4$ and $d = 2$.*

The following description of the geometric permutations in the case of unit radius will be used in Section 5:

THEOREM 4.2. *Two geometric permutations of n disjoint unit balls in $\mathbb{R}^d$, with $n \geq 9$ or $n \geq 4$ and $d = 2$, differ by switching two adjacent elements.*

These bounds were obtained through essentially three techniques we now review: *separation sets*, *switch pairs* and *incompatible pairs*.

4.1. Separation sets. Let $\mathcal{C}$ be a collection of disjoint convex sets in $\mathbb{R}^d$. A *separation set* $\mathcal{H}$ for $\mathcal{C}$ is a set of hyperplanes such that any two members in $\mathcal{C}$ can be separated by a hyperplane parallel to some element in $\mathcal{H}$. An oriented transversal ℓ to two disjoint convex sets C_1 and C_2 meets C_1 first if and only if for some hyperplane Π separating C_1 and C_2, ℓ meets the halfspace containing C_1 first; in other words, if Γ_Π denotes the hypersphere of directions of Π, it depends on which side of Γ_Π $\overrightarrow{\ell}$ lies. Thus, the geometric permutation realized by a transversal to $\mathcal{C}$ depends only on the cell of the arrangement on $\mathbb{S}^{d-1}$ of the hyperspheres of directions associated with the members of $\mathcal{H}$ that contains its direction. As a consequence, the number of geometric permutations of $\mathcal{C}$ is bounded by the complexity of that arrangement, that is $O\left(|\mathcal{H}|^{d-1}\right)$, and n disjoint compact convex objects have $O(n^{2d-2})$ geometric permutations [78].

Upper bound for balls. Collections of disjoint balls admit small separation sets [73, Theorem 4.1]. The argument goes as follows. Let $\mathcal{C} = \{B_1, \ldots, B_n\}$ be a collection of disjoint balls in $\mathbb{R}^d$. Cover the sphere of directions $\mathbb{S}^{d-1}$ by spherical caps $C_1, \ldots, C_k$ of given opening angle α. For any $1 \leq i \leq n$ and $1 \leq j \leq k$ let $\Gamma_{i,j}$ denote the cone with apex the center of B_i induced by cap C_j and $h_{i,j}$ a hyperplane separating B_i from the closest ball with larger radius and having its center in $\Gamma_{i,j}$, if any; specifically, $h_{i,j}$ is chosen tangent to B_i and normal to the line through the centers of the two separated balls (see Figure 5).

For α smaller than $\sin^{-1}\left(\frac{\sqrt{3}-1}{2}\right)$, the collection

$$\{h_{i,j} | 1 \leq i \leq n, 1 \leq j \leq k\}$$

separates any two balls in $\mathcal{C}$. As a consequence, any collection of n disjoint balls admits a separation set of size $O(n)$, and has $O(n^{d-1})$ geometric permutations.

Lower bound for balls. The previous upper bound is asymptotically tight [73, Theorem 4.5]. Consider n hyperplanes $H_1, \ldots, H_n$ in $\mathbb{R}^d$ going through the origin, no d of them containing a line, and let S_i denote the set of directions parallel to H_i. Let $\epsilon > 0$ be small enough such that any cell in the arrangement $\mathcal{A}$ of $\{S_1, \ldots, S_n\}$ on $\mathbb{S}^{d-1}$ contains a point at distance at least ϵ from every S_i. For $\delta > 0$, let $(B_i^1(\delta), B_i^2(\delta))$ be two balls centered on the perpendicular to H_i through the origin, at distance δ from the origin and separated by H_i; $B_i^1(\delta)$ and $B_i^2(\delta)$ have equal radius, chosen such that a line through the origin intersects them if and only if it makes an angle at least ϵ with H_i. The construction consists of a pair $(B_i^1(\delta_i), B_i^2(\delta_i))$ for

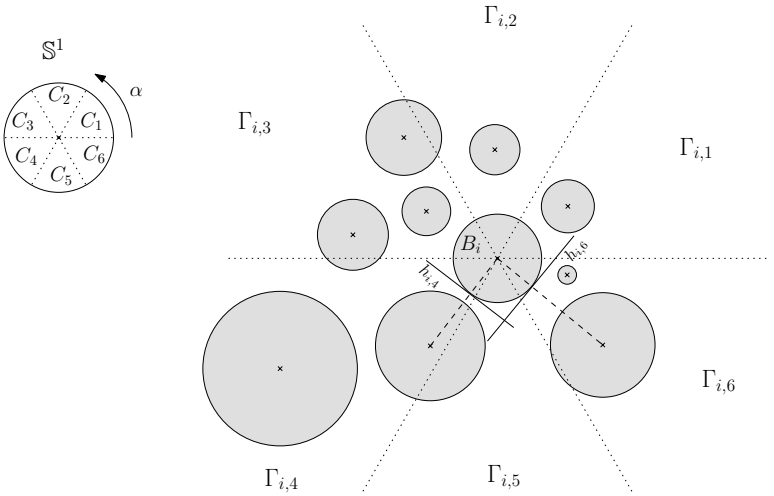


FIG. 5. *Construction of a small separation set of a collection of balls ($d = 2$, $k = 6$, $\alpha = \frac{\pi}{3}$).*

$i = 1, \ldots, n$, where $\delta_1 = 1$ and δ_{i+1} is chosen larger than the diameter of $\bigcup_{1 \le t \le i} \left(B_t^1(\delta_t) \cup B_t^2(\delta_t)\right)$. By construction, any line through the origin with direction at least ϵ away from each of the S_i intersects all the balls. There are as many classes of such lines as cells in $\mathcal{A}$, that is $\Omega(n^{d-1})$, and two lines with directions in different cells realize different orderings of the balls.

4.2. Switch pairs. Let $\mathcal{C}$ be a family of n disjoint balls in $\mathbb{R}^d$ with radius disparity at most γ that admits some transversal. Assume, w.l.o.g. that the radius of the smallest ball in $\mathcal{C}$ is 1. For all asymptotic estimates we assume that d is constant and $n \gg \gamma^{d-1}$.

For n large enough, the transversals to $\mathcal{C}$ are nearly parallel. Specifically, a volume argument shows that the diameter of the set of centers of balls in $\mathcal{C}$ is $\Omega\left(\frac{n}{\gamma^{d-1}}\right)$; the angle between two transversals to $\mathcal{C}$ is then $O\left(\frac{\gamma^d}{n}\right)$ as the distance between them is at most 2γ along segments of length $\Omega\left(\frac{n}{\gamma^{d-1}}\right)$. We say that two transversals are oriented *consistently* if the angle between their direction vectors is less than $\pi/2$, that is, close to 0 when the transversals are nearly parallel. A *switch pair* for $\mathcal{C}$ is a pair of balls intersected in different orders by two transversals to $\mathcal{C}$ oriented consistently[14]. Switch pairs were investigated first for congruent disks in the plane [72, 73], then for balls in higher dimensions, both for the unit radius case [20, 48, 53] and the bounded radius disparity case [48, 81].

[14] A similar, but more general, notion is investigated by Asinowski et al. [9].

4.2.1. Properties of a switch pair. For n sufficiently large, a ball participates in at most one switch pair and two balls in a switch pair appear consecutively in any geometric permutation of $\mathcal{C}$ ([53, Lemmas 2.8 and 2.9], [48, Lemmas 8 and 9] and [81, Lemmas 2.10 and 2.11]). These properties follow from simple geometric considerations:

- the distance between two balls in a switch pair is $O\left(\frac{\gamma^{2d+1}}{n^2}\right)$ ([53, Lemmas 2.6 and 2.7], [48, Lemma 6] and [81, Lemma 2.7]),
- the line through the centers of the balls in a switch pair makes angle $\frac{\pi}{2} - O\left(\frac{\gamma^d}{n}\right)$ with any transversal to $\mathcal{C}$ ([53, Lemmas 2.6 and 2.7], [48, Lemma 6] and [81, Lemma 2.8]).

Consequently, to bound the number of geometric permutations of $\mathcal{C}$ it suffices to bound its number of switch pairs, as k switch pairs allow at most 2^k geometric permutations.

4.2.2. Number of switch pairs. The number of switch pairs can be bounded via considerations on the distances between their inner centers of similitude. Let Δ be the line through the centers of the two balls furthest apart. The disjointedness of the balls and the upper bound on the distance between two members of a switch pair imply that the projection on Δ of the inner centers of similitude of two switch pairs are distance at least $\sqrt{2}-o(1)$ and at most $2\gamma + o(1)$ apart ([81, Theorem 4.3], [53, Lemma 3.2] and [48, Lemma 16]). An upper bound of $1 + \lfloor\sqrt{2}\gamma\rfloor$ on the number of switch pairs follows when n is large enough. In the planar case, a different argument based on incompatible pairs (see Section 4.3) yields the same result [81].

4.2.3. Hamming distance between geometric permutations. The previous bound on the number of switch pairs yields that sufficiently large collections of disjoint balls with radius disparity γ have at most $2^{1+\lfloor\sqrt{2}\gamma\rfloor}$ geometric permutations. Number the m switch pairs of $\mathcal{C}$ and assign to every geometric permutation of $\mathcal{C}$ a vector in $\{0,1\}^m$ depending on the ordering in which each pair is traversed. If two geometric permutations differ by the switching of k pairs then the radius disparity of the balls is at least $2^{\lceil\frac{k}{2}\rceil-1}$ [81]. Thus, the number of elements that differ, i.e. the *Hamming distance*, between the vectors of two geometric permutations is bounded by $2(1+\lfloor\log\gamma\rfloor)$. The size of a subset of $\{0,1\}^m$ with diameter at most 2δ under the Hamming distance is $O\left(\frac{(4m)^\delta}{\delta!}\right)$. Therefore, disjoint balls with radius disparity at most γ have $O\left(\gamma^{\log\gamma}\right)$ geometric permutations.

4.2.4. The case of unit balls. The previous result implies that sufficiently large collections of disjoint unit balls have at most 4 geometric permutations. This bound was reduced to 2 by ad hoc techniques.

In the plane. If $\mathcal{C}$ admits 3 geometric permutations it has at least two switch pairs, say (A, B) and (C, D). Up to symmetries we can then assume that the four disks admit the three geometric permutations $ABCD$,

TABLE 1
The 12 geometric permutations on $1, \ldots, 4$.

α	1234	2143	1432	4123
β	1243	4312	1342	3124
γ	1324	3142	1423	4132

$BACD$, and $ABDC$. Thus, the cells of the Voronoi diagram of the centers of these disks, where each cell inherits the label of the disk it contains, also admit these three geometric permutations. An elementary case-study of the configurations of four points in the plane shows that this is impossible [73]. Thus, any sufficiently large family of disjoint unit disks in $\mathbb{R}^2$ has at most one switch pair and at most 2 geometric permutations.

In higher dimensions. A refined analysis shows that the distance between the inner centers of similitude of two switch pairs consisting of unit balls is at most $1 + o(1)$ [20, Lemma 4]. Since this contradicts the lower-bound of $\sqrt{2} - o(1)$ previously obtained for the same distance, it proves that sufficiently large collections of disjoint unit balls have at most one switch pair and 2 geometric permutations.

4.3. Incompatible pairs. An efficient way to bound the size of a set of permutations is to show that certain patterns cannot occur. Given two geometric permutations gp_1 and gp_2 of $\mathcal{C}$ and two permutations on k elements p_1 and p_2, we say that (p_1, p_2) is a sub-pattern of (gp_1, gp_2) if the restriction of (gp_1, gp_2) to some k objects in $\mathcal{C}$ yields two permutations that are equal, up to relabelling and reversing, to (p_1, p_2). Showing that certain pairs of permutations on four elements cannot occur as sub-patterns of pairs of geometric permutations of disjoint unit balls led to bounds that are tight in the plane [7] and almost tight in higher dimensions [21]. In particular, these bounds also apply to small families, unlike those obtained by studying switch pairs. The use of incompatible pairs for studying geometric permutations can be traced back to Katchalski et al. [51, Section 5], although they use a different presentation.

4.3.1. Families with incompatible pairs. To bound the number of geometric permutations of disjoint unit balls, the incompatible pairs investigated are of size 4 (for situations where larger families were considered see eg. [8, 10]). The pairs of geometric permutations considered for disjoint unit balls are

$$a = (1234, 2143), \qquad b = (1234, 1432),$$
$$c = (1234, 1423), \qquad d = (1234, 3142),$$

and Table 1 summarizes all 12 geometric permutations on $1, \ldots, 4$, divided into three rows. Any pair of permutations in row α is equal, up to relabelling and reversing, to one of a or b and row β or γ can be obtained from

α by adequate relabelling. Therefore, any quadruple of objects for which a and b are incompatible has at most three geometric permutations, one from each row (the same holds if b is replaced by $(1234, 1423)$ [51]). Given three different geometric permutations $\sigma_1, \ldots, \sigma_3$ of $n \geq 4$ objects for which a and b are incompatible, there are always 3 objects to which the restrictions of the σ_i differ [21, Lemma 1]; from there, one can prove that any family of n objects for which a and b are incompatible has at most three geometric permutations [21, Lemma 2]. Note that pairs other than a and b can be used equivalently, for instance a and $(1234, 4123)$ [7].

Similar arguments yield that any family of n objects for which the pairs $a, \ldots, d$ are incompatible has at most two geometric permutations that differ only by the swapping of two adjacent elements [21, Lemma 3]. Proving Theorems 4.1 and 4.2 thus reduces to showing that the pairs a to d are incompatible.

4.3.2. The planar case. Consider two intersecting transversals to four disks, and mark on each line one point from each disk. Different situations arise depending on which half-line each point belongs (there are 29 such configurations). A careful analysis of these situations shows that for families of disjoint unit disks the pairs a and b [51, Lemmas 1–3] and c and d [7][15] are incompatible[16]. A different proof, avoiding the discussion of the 29 configurations, was given later by the same authors [9].

4.3.3. Higher dimensions. A proof that pair a is incompatible for disjoint unit balls in $\mathbb{R}^d$ can be obtained through elementary, although tedious, geometric observations [21, Section 4]; this analysis essentially refines the earlier proof that sufficiently large collections of disjoint unit balls have at most one switch pair [20], another way to formulate the incompatibility of a. The proof of incompatibility of b, c and d rests on the following crucial observation:

LEMMA 4.1 ([21], Lemma 7). *Let $\overrightarrow{v}$ be a direction of a transversal to 3 disjoint unit balls in $\mathbb{R}^d$ and $\overrightarrow{u}$ the vector from the center of the first to the center of the last ball met by that transversal. Then $\angle(\overrightarrow{v}, \overrightarrow{u}) < \pi/4$.*

To see that pair b is incompatible, let $\overrightarrow{v_1}$ and $\overrightarrow{v_2}$ be two directions of transversals intersecting four disjoint unit balls in, respectively, the orders 1234 and 1432. Let c_i denote the center of ball i. By Lemma 4.1 we have that $\angle(\overrightarrow{v_1}, -\overrightarrow{v_2}) < \pi/2$ since both $\overrightarrow{v_1}$ and $-\overrightarrow{v_2}$ make an angle less than $\pi/4$ with $\overrightarrow{c_2c_4}$. Also, $\angle(\overrightarrow{v_1}, \overrightarrow{v_2}) < \pi/2$ as both $\overrightarrow{v_1}$ and $\overrightarrow{v_2}$ make an angle less than $\pi/4$ with $\overrightarrow{c_1c_3}$, and we get a contradiction.

A packing argument ([21, Lemma 6]) shows that the intersection of the solids bounded by two cylinders of radius 1 whose axis make angle at least $\pi/4$ contains at most 8 points with smallest inter-point distance at

[15] The same result was obtained independantly by A. Holmsen in his master's thesis.

[16] For families of disjoint translates of a convex set, pairs a and b remain incompatible [51] but pairs c and d cannot be *both* incompatible: indeed, there exist arbitrarily large such families with three geometric permutations.

least 2. Thus, two transversals to any collection of $n \geq 9$ disjoint unit balls in $\mathbb{R}^d$ make an angle of less than $\pi/4$. Now, let $\overrightarrow{v}$ and $\overrightarrow{v'}$ be the direction vectors of transversals to $n \geq 9$ disjoint unit balls that realize, respectively, the permutations 1234 and one of 1423 or 3142 on some subset of four balls. Because there are $n \geq 9$ balls,

$$\angle(\overrightarrow{v}, \overrightarrow{v'}) < \frac{\pi}{4}$$

and Lemma 4.1 implies that the angle between $\overrightarrow{v}$ and $\overrightarrow{c_2c_4}$ is at most $\frac{\pi}{4}$. Consequently, the angle between $\overrightarrow{v'}$ and $\overrightarrow{c_2c_4}$ is less than $\frac{\pi}{2}$ and the second line should meet ball 2 before ball 4, a contradiction. Thus pairs c and d are incompatible for any family of $n \geq 9$ disjoint unit balls in $\mathbb{R}^d$. This proves Theorem 4.1 and Theorem 4.2 for $d \geq 3$.

5. Pinning, Hadwiger and Helly numbers. The Helly-type theorems for transversals to disjoint balls essentially generalize two landmark results in geometric transversal theory due to Hadwiger and Danzer.

Hadwiger's transversal theorem states that n disjoint[17] convex sets in the plane have a transversal if any 3 have a transversal consistent with some global ordering of the family [41]. The bound on the Hadwiger number shows that this theorem generalizes to disjoint balls in arbitrary dimension, a situation that is remarkable as it is not the case for disjoint translates of a convex set, not even in $\mathbb{R}^3$ [46].

Danzer proved that n disjoint unit disks[18] in the plane have a transversal if any 5 do [23], and conjectured that a similar result holds in higher dimensions. The bound on the Helly number for disjoint unit balls is the positive answer to this question.

5.1. Relationship between the pinning and Hadwiger numbers. In the plane, the pinning and Hadwiger numbers are the same, namely 3. In higher dimensions, the convexity of the cone of directions (Theorem 3.1) implies:

THEOREM 5.1. *If p_d and h_d denote respectively the pinning and Hadwiger numbers of collections and sequences of disjoint balls in $\mathbb{R}^d$ then $h_d \leq p_d + 1$.*

Proof. Let $\mathcal{C}$ be a sequence of at least $n \geq p_d + 2$ disjoint balls and assume that every subsequence of size p_d+1 has an order-respecting transversal. Shrink continuously all balls by, for instance, multiplying all radii by some parameter t ranging from 1 down to 0, until some subsequence X of size $p_d + 1$ is about to lose its last order-respecting transversal. By Theorem 3.1, at that position $\mathcal{K}(X)$ is a single point and X has a unique

[17] This assumption can be dropped [77].

[18] Grünbaum proved the same statement for unit axis-parallel squares [37], and conjectured that it holds for collections of disjoint translates of a convex set, a conjecture proven 30 years later by Tverberg [75].

order-respecting line transversal ℓ. Since (X, ℓ) is a pinning configuration, there exists a subset $Y \subset X$ of size at most p_d such that Y pins ℓ. Given any $Z \in (\mathcal{C} \setminus Y)$, the subsequence $Y \cup \{Z\}$ has size at most $p_d + 1$ and thus has some order-respecting line transversal ℓ_Z. Since (Y, ℓ) is a pinning configuration, Y admits no order-respecting transversal other than ℓ, and thus $\ell_Z = \ell$ and ℓ intersects Z. It follows that $\mathcal{C}$ has a line transversal, and $h_d \leq p_d + 1$. □

REMARK 5.1. The same proof yields that if every subsequence of size $p_d + 2$ has an order-respecting transversal then $\mathcal{C}$ has an *order-respecting* transversal.

5.2. Bounds on the pinning and Hadwiger numbers. The current bounds on the pinning and Hadwiger numbers, given by Theorem 5.3, grow linearly with the dimension [19]. We sketch the proof of these bounds, after giving a much simpler argument that yields a bound quadratic in the dimension [3, 40, 45].

5.2.1. A simple quadratic bound. The pinning and Hadwiger numbers can be bounded by applying Helly's theorem successively on $\mathbb{S}^d$ and on the projections along certain directions, an argument already used by Vincensini [76].

LEMMA 5.1. *The pinning and Hadwiger numbers of disjoint balls in* $\mathbb{R}^d$ *are bounded from above by* $d(d+1)$.

Proof. Let $\mathcal{C}$ be a sequence of disjoint balls in $\mathbb{R}^d$ and $\binom{\mathcal{C}}{d}$ the set of its subsequences of length d. As argued in the proof of Lemma 3.1,

$$\mathcal{K}(\mathcal{C}) = \bigcap_{X \in \binom{\mathcal{C}}{d}} \mathcal{K}(X),$$

and Theorem 3.1 yields that for any subsequence X the set $\mathcal{K}(X)$ is strictly convex.

The spherical Helly theorem on $\mathbb{S}^{d-1}$ (Theorem 2.2) implies that $\mathcal{K}(\mathcal{C})$ is nonempty if and only if for any $d+1$ elements $X_1, \ldots, X_{d+1} \in \binom{\mathcal{C}}{d}$ the intersection $\bigcap_{1 \leq i \leq d+1} \mathcal{K}(X_i)$ is nonempty. In other words, $\mathcal{C}$ has a transversal[19] if and only if any subsequence of length at most $d(d+1)$ has an order-respecting line transversal. This proves the statement for the Hadwiger number.

Similarly, $\mathcal{K}^o(\mathcal{C})$ is the intersection of the $\mathcal{K}^o(X)$ for $X \in \binom{\mathcal{C}}{d}$. Thus, if $\mathcal{C}$ pins some order-respecting transversal ℓ, the same arguments yield that $\mathcal{K}^o(\mathcal{C})$ is empty if and only if $\mathcal{K}^o(X)$ is empty for some subsequence $X \subset \mathcal{C}$ of length at most $d(d+1)$. Since $\mathcal{K}(\mathcal{C}) \subset \mathcal{K}(X)$ we deduce that $\mathcal{K}(X)$ is a single point, and X pins ℓ as well. This proves the statement for the pinning number. □

REMARK 5.2. If the balls are unit this bound becomes d^2: by Lemma 4.1, the cone of directions of any sequence of $n \geq 3$ balls has

[19] In fact, $\mathcal{C}$ has an order-respecting transversal.

opening angle at most $\frac{\pi}{4}$, so we can apply Helly's theorem in $\mathbb{R}^{d-1}$ instead of $\mathbb{S}^{d-1}$ (Theorem 2.3 instead of Theorem 2.2) in the previous proof.

5.3. A linear bound. For thinly distributed balls, Grünbaum [38] obtained a linear bound on the Hadwiger number by applying Helly's topological theorem directly to the set of line transversals. More generally[20], we have:

THEOREM 5.2. *Let $\mathcal{U}_d$ be the set of all collections of balls in $\mathbb{R}^d$ admitting a separation set of size 1. The pinning, Hadwiger and Helly numbers of $\mathcal{U}_d$ are bounded from above by $2d-1$.*

Proof. Let $\mathcal{C} = \{B_1, \ldots, B_n\}$ be a sequence of balls in $\mathbb{R}^d$ with separation set $\{H\}$ (in particular the balls are pairwise disjoint). Let ϵ denote the minimal angle any transversal to *two* balls in $\mathcal{C}$ makes with H, and let $T(X)$ denote the set of transversals to a subsequence $X \subset \mathcal{C}$ making an angle at least ϵ with H. Parameterizing lines by their intercept in two translated copies of H recasts the $T(B_i)$ as contractible subsets of $\mathbb{R}^{2d-2}$. Thus, to apply Helly's topological theorem it suffices to prove:

$$\forall A_1, \ldots, A_{2d-2} \in \mathcal{C}, \qquad \bigcap_{1 \le i \le 2d-2} T(A_i) \text{ is a homology cell.}$$

Since $\mathcal{C}$ has a separation set of size one, any subsequence of $\mathcal{C}$ has at most one geometric permutation. Therefore, for $X, Y \subset \mathcal{C}$ we have that

$$T(X) \cap T(Y) = T(X \cup Y)$$

and the above condition follows from Theorem 3.2. Therefore, the Helly number for thinly distributed balls in $\mathbb{R}^d$ is at most $2d-1$. Because all subsequences have a unique geometric permutation, the bound on the Hadwiger number follows. Since a transversal is isolated if the open balls have no transversal in the same order, the bound on the pinning number also follows. □

Compatible directions. The same idea can be applied to more general families of balls by restricting the set of possible directions of transversals so that any subsequence of $\mathcal{C}$ has only one geometric permutation [19]. Specifically, call a direction $\overrightarrow{u}$ *compatible* with a sequence $\mathcal{C}$ if

$$\forall A \prec B \text{ in } \mathcal{C}, \quad \langle \overrightarrow{u}, \overrightarrow{ab} \rangle > 0,$$

where a and b denote the respective centers of A and B; by extension, we say that a transversal to $X \subset \mathcal{C}$ is *compatible* if its direction is. The directions of compatible transversals to a subsequence $X \subset \mathcal{C}$ are the intersection of $\mathcal{K}(X)$ with a polytope in $\mathbb{S}^{d-1}$, and thus strongly convex; the same proof as in Theorem 3.2 yields that the set of compatible transversals to

[20] Grünbaum's proof exploits the fact that the distance condition that characterizes thinly distributed balls guarantees that the family has a separation set of size 1.

$X \subset \mathcal{C}$ is contractible. Also, a compatible transversal to $X, Y \subset \mathcal{C}$ is order-respecting on $X \cup Y$. As a consequence, the proof of Theorem 5.2 yields [19, Lemma 15]:

LEMMA 5.2. *If $\mathcal{C}$ is a sequence of disjoint open balls such that any subset of size $2d-1$ has a transversal compatible with $\mathcal{C}$, then $\mathcal{C}$ has a compatible transversal.*

We can now bound the pinning and Hadwiger numbers [19, Proposition 13]:

THEOREM 5.3. *The pinning and Hadwiger numbers for disjoint balls in $\mathbb{R}^d$ are bounded from above by, respectively, $2d-1$ and $2d$.*

Proof. Let $\mathcal{C}$ be a sequence of disjoint balls in $\mathbb{R}^d$ that pins an order-respecting transversal ℓ. From Corollary 3.2 we get that the open balls in $\mathcal{C}$ have no compatible transversal, and so Lemma 5.2 yields that some subsequence $X \subset \mathcal{C}$ of size at most $2d-1$ has no transversal compatible with $\mathcal{C}$. Since $\mathcal{K}(X)$ is convex and the set of compatible directions, which is open, intersects $\mathcal{K}(X)$ but not $\mathcal{K}^o(X)$, $\mathcal{K}(X)$ has empty interior and X pins ℓ. Thus, the pinning number is at most $2d-1$ and Theorem 5.1 bounds the Hadwiger number by $2d$. □

5.4. Helly numbers. A family of examples by Danzer [23] (see also [44, Figure 3]) shows that the Helly number of disjoint balls is already unbounded in dimension 2. It can, still, be bounded under additional assumptions, e.g. for thinly distributed balls (Theorem 5.2). The case of congruent balls received particular attention and the bounds can be summarized as follows:

THEOREM 5.4. *The Helly number of families of disjoint unit balls in $\mathbb{R}^d$ is 5 for $d = 2$ and at most $4d-1$ for $d \geq 3$.*

We describe in Sections 5.4.1 and 5.4.2 the two techniques used to obtain such bounds. We show in Section 5.4.3 that the requirement that the balls be unit can be replaced by a bound on the number of geometric permutations of all subfamilies (Theorem 5.6). This implies, for instance, that the Helly number of a family of balls with radius disparity at most γ can be bounded by a function of d and γ (Corollary 5.1).

5.4.1. Designing an ordering. Holmsen et al. [45] used the earlier analysis of switch pairs [20, 48, 53, 81] to bound the Helly number of disjoint unit balls in 3 dimensions by 22.

Let δ denote the smallest diameter of a set of centers of 31 disjoint unit balls in $\mathbb{R}^3$ and $\mathcal{C} = \{B_1, \ldots, B_n\}$ a collection of at least 31 disjoint unit balls in $\mathbb{R}^3$. Assume that the centers of B_1 and B_n are the furthest apart and let $\mathcal{T}$ denote the set of transversals to these two balls, oriented from B_1 to B_n. Say that (B_i, B_j) is a *switch pair* if there are transversals in $\mathcal{T}$ that meet these two balls in distinct orders[21]. A result similar to

[21] Note that this definition slightly differs from that used in Section 4.

Theorem 4.2 applies to families with less than 9 balls provided the centers are sufficiently spread out [45, Theorem 3]:

LEMMA 5.3. *Any family of disjoint unit balls in $\mathbb{R}^3$ whose set of centers has diameter at least δ has at most two switch pairs; the balls of a switch pair appear consecutively in any geometric permutation of the family.*

We can then bound the Helly number as follows [45, Theorem 1]:

PROPOSITION 5.1. *The Helly number of collections of at least 31 disjoint unit balls in $\mathbb{R}^3$ is at most 22.*

Proof. Assume that every subset of $\mathcal{C}$ of size at most 22 has a transversal. We discuss the case where $\mathcal{C}$ has two switch pairs P_1 and P_2 (if $\mathcal{C}$ has one or no switch pair the proof is similar). Lemma 5.3 implies that there exists an ordering $\prec'$ on $\mathcal{C}' = \mathcal{C} \setminus (P_1 \cup P_2)$ such that any transversal in $\mathcal{T}$ to a subset of $\mathcal{C}'$ respects $\prec'$. Since the balls in each switch pair are consecutive, there are only 4 possible extensions of $\prec'$ into an ordering of $\mathcal{C}$, say $\prec_1, \ldots, \prec_4$. Assume that for each $i = 1, \ldots, 4$ there is a quadruple $Q_i \subset \mathcal{C}'$ such that $Q_i \cup P_1 \cup P_2$ has no transversal in $\mathcal{T}$ respecting $\prec_i$. Then the at most 22 balls of the subset

$$\left(\bigcup_{1 \le i \le 4} Q_i \right) \bigcup P_1 \bigcup P_2 \bigcup \{B_1, B_n\}$$

have no common transversal, which contradicts the assumption. Consequently, some extension $\prec_i$ of $\prec$ is such that for any quadruple $Q \subset \mathcal{C}'$ the balls in $Q \cup P_1 \cup P_2$ have a transversal in $\mathcal{T}$ respecting $\prec_i$. It follows that every 6-tuple in $\mathcal{C}$ has a transversal respecting $\prec_i$, and since the Hadwiger number of disjoint balls in $\mathbb{R}^3$ is at most 6 (by Theorem 5.3), we get that $\mathcal{C}$ has a transversal. □

REMARK 5.3. This approach extends naturally to higher dimensions, resulting on a bound of $4h_d - 2$ for the Helly number of sufficiently large families of disjoint unit balls, where h_d is the corresponding Hadwiger number. The threshold above which a family of balls is "sufficiently large" increases with the dimension.

REMARK 5.4. Holmsen et al. [45] used a bound of 12 on the Hadwiger number of disjoint balls in $\mathbb{R}^3$, thus obtaining a bound of 46 on the Helly number. Their theorem thus omits the assumption that the family be large enough.

5.4.2. The homotopy method. The technique used to bound the Hadwiger number in terms of the pinning number in Theorem 5.1 can also be used to bound the pinning number of disjoint unit balls [19, Theorem 2]:

THEOREM 5.5. *The Helly number for disjoint unit balls in $\mathbb{R}^d$, $d \ge 2$, is bounded from above by $2p_d + 1$, where p_d denotes the pinning number, for disjoint balls in $\mathbb{R}^d$.*

Proof. Let $\mathcal{C}$ be a collection of disjoint unit balls in $\mathbb{R}^d$ such that any subset of $2p_d + 1$ balls has a transversal, where p_d denotes the pinning

number for disjoint (unit) balls in $\mathbb{R}^d$. Shrink uniformly the balls in $\mathcal{C}$ until the first subset of $2p_d+1$ balls, say $\mathcal{F}$, is about to lose its last order-respecting transversal. Any subset of $\mathcal{F}$ of size at least $|\mathcal{F}|-2$ has at most two geometric permutations differing by the switching of two consecutive balls [22]. In the rest of this proof all balls are considered shrunk.

We first argue that we can assume that $\mathcal{F}$ has a unique transversal ℓ. Otherwise, $\mathcal{F}$ has only isolated transversals, each one corresponding to a distinct geometric permutation. Theorem 4.2 yields that there are 2 such lines, say ℓ_1 and ℓ_2. Each ℓ_i can be pinned by p_d balls from $\mathcal{F}$, so some subset $\mathcal{F}' \subset \mathcal{F}$ of size $|\mathcal{F}|-1$ suffices to pin both of them and Theorem 4.2 implies that this subset has no other transversal. Consequently, every ball in $\mathcal{C} \setminus \mathcal{F}'$ meets one of the ℓ_i. If all such balls meet ℓ_2, it is a transversal to $\mathcal{C}$ and we are done; if some ball A misses ℓ_2 then $\mathcal{F}' \cup \{A\}$ is a subset of size at most $|\mathcal{F}|$ with $\ell = \ell_1$ as unique transversal.

Next, we argue that some proper subset $\mathcal{F}'$ of $\mathcal{F}$ has no other transversal than ℓ. Otherwise, let $\mathcal{G}$ be a subset of size p_d that pins ℓ and for $Z \in \mathcal{F} \setminus \mathcal{G}$ denote by ℓ_Z a transversal to $\mathcal{F} \setminus \{Z\}$ other than ℓ. Since $\mathcal{G}$ pins ℓ, the orderings $\prec_\ell$ and $\prec_{\ell_Z}$ differ on $\mathcal{G}$ and thus on $\mathcal{F} \setminus \{Z\}$ and, by Theorem 4.2, they differ by the swapping of two balls X_Z and Y_Z. Since $\prec_\ell$ and $\prec_{\ell_Z}$ already differ on $\mathcal{G}$, we have that $X_Z, Y_Z \in \mathcal{G}$. For $A, B \in \mathcal{F} \setminus \mathcal{G}$, the set $\mathcal{F} \setminus \{A, B\}$ has three transversals (ℓ, ℓ_A and ℓ_B) but, by Theorem 4.2, at most two geometric permutations. Since $\prec_\ell$ and $\prec_{\ell_Z}$ disagree on $\mathcal{G}$, we thus get that ℓ_A and ℓ_B induce the same geometric permutation on $\mathcal{F} \setminus \{A, B\}$ for any $A, B \in \mathcal{F} \setminus \mathcal{G}$. It follows that X_Z and Y_Z are independent of the choice of Z; call these two balls X and Y and let $\prec$ be the ordering on $\mathcal{F}$ obtained by swapping X and Y in $\prec_\ell$. Since $\mathcal{F}$ has no transversal in the order $\prec$, Remark 5.1 implies that some subset $\mathcal{H} \subset \mathcal{F}$ of size p_d+2 has no transversal respecting that order. Thus, $\mathcal{F}' = \mathcal{G} \cup \mathcal{H}$ has no other transversal than ℓ; the balls X and Y both belong to $\mathcal{H}$ as otherwise $\prec$ and $\prec_\ell$ are equivalent, so $\mathcal{F}'$ has size at most $2p_d$ and is a proper subset of $\mathcal{F}$.

Let X be some ball in $\mathcal{C}$. Since $\mathcal{F}' \cup \{X\}$ has some transversal and ℓ is the only transversal to $\mathcal{F}'$, it follows that ℓ intersects X. Thus, $\mathcal{C}$ has a transversal. □

Note that this bound is not tight in the two-dimensional case [23].

5.4.3. Using Helly's theorem for unions of sets. Using Matousek's generalization of Helly's topological theorem (see Theorem 2.6), we can replace the constraint on the radii radii by considerations on numbers of geometric permutations:

THEOREM 5.6. *For any $d \geq 2$ and $k \geq 1$ there exists a number $h^\star(d,k)$ with the following property. Let $\mathcal{C}$ be a finite family of disjoint open balls in $\mathbb{R}^d$ such that any sub-family of $\mathcal{C}$ has at most k geometric permutations. Then $\mathcal{C}$ has a line transversal if and only if every sub-family of size at most $h^\star(d,k)$ has a line transversal.*

[22] Theorem 4.2 applies as $p_2 = 3$ and $p_d \geq 5$ for $d \geq 3$ (see Section 5.5).

Proof. Let $\mathcal{C} = \{B_1, \ldots, B_n\}$ be a family of disjoint balls in $\mathbb{R}^d$ and T_i the set of oriented line transversals to B_i. By Theorem 3.2, the intersection of any number of T_i consists of at most k contractible components. The T_i are subsets of the Grassmaniann $G_{2,2d-2}$, which naturally embeds in $\mathbb{P}^{2d-1}$; to apply Theorem 2.6, we embed (part of) the T_i in $\mathbb{R}^{2d-2}$.

We handle this technicality as follows. Let Π and Π' be two parallel planes and T_i^* denote the set of oriented line transversals to ball B_i that are not parallel to Π. By parameterizing all lines not parallel to Π using their intersections with Π and Π', we recast the T_i^* as subsets of $\mathbb{R}^{2d-2}$. The directions of lines in T_i^* are exactly the directions of lines in T_i minus a great hypersphere. Similarly, the directions of an intersection of T_i^* consists in the difference of at most k convex sets and a great hypersphere, which is at most $2k$ convex sets. This implies that the intersection of any T_i^* has at most $2k$ connected components, and Theorem 2.6 applies. With $h^\star(d,k) = h(2d-2, 2k)$, we thus get that if every subset of $\mathcal{C}$ of size $h^\star(d,k)$ has a line transversal not parallel to Π then $\mathcal{C}$ has a line transversal. Now, observe that if a subset has a strict line transversal, then its cone of directions has nonempty interior and it must have a line transversal not parallel to Π. □

By Theorem 4.1, this applies to case of balls of bounded radius disparity immediately:

COROLLARY 5.1. *The Helly number of a family of disjoint balls in $\mathbb{R}^d$ with radius disparity at most γ can be bounded by a function of d and γ.*

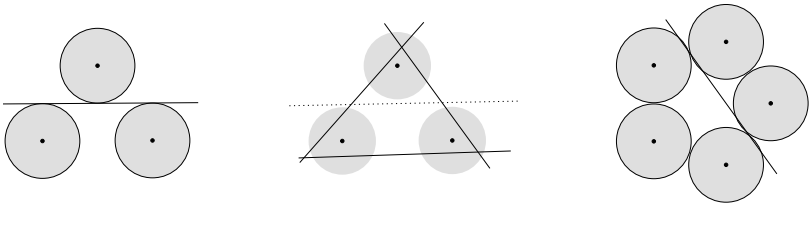
REMARK 5.5. Unfolding the same approach using Amenta's generalization of Helly's theorem (see Corollary 2.1) requires to control the intersection of a set $\mathcal{F}$ of directions of transversals to d-tuples of balls. If $\mathcal{F}$ is the set of all d-tuples of a family of balls, then this intersection consists of at most k disjoint convex sets on the sphere $\mathbb{S}^{d-1}$. If $\mathcal{F}$ consists of some but not all d-tuples of a family of balls, the components of this intersection are still convex, but it is not clear what their number is.

5.5. Lower bounds. This section discusses the few lower bounds known for the pinning, Hadwiger and Helly numbers. First, we observe that these numbers are monotone in the dimension:

THEOREM 5.7. *The pinning, Hadwiger and Helly numbers of disjoint balls in $\mathbb{R}^d$ are nondecreasing in d.*

Proof. Let $\mathcal{C}$ be a collection of balls in $\mathbb{R}^d$ such that all centers lie in some k-flat Π. If ℓ is a transversal to $\mathcal{C}$ then so is the orthogonal projection of ℓ on Π, as the orthogonal projection reduces the distance to the balls' centers. As a consequence, any lower-bound example for the pinning, Hadwiger and Helly numbers in $\mathbb{R}^k$ can be embedded in $\mathbb{R}^d$ for $d \geq k$ while retaining its transversal properties, and these numbers are nondecreasing with d. □

Two-dimensional examples. In the plane, the pinning and Hadwiger numbers of disjoint disks are at most 3; these bounds are easily seen to be



(a) (b) (c)

FIG. 6. *Lower bounds in the plane for the pinning (a), Hadwiger (b) and Helly (c) numbers.*

tight (c.f. Figure 6(a) and (b)). Also, the Helly number of disjoint unit disks is exactly 5, as follows from the example of five unit disks centered at the vertices of a regular pentagon, depicted by Figure 6(c) (see [23, 44] for a more detailed description).

Higher dimensions. The upper bound for the pinning number of families of disjoint balls can be shown to be tight [18]:

THEOREM 5.8. *The pinning number of disjoint balls in* $\mathbb{R}^d$ *is exactly* $2d-1$.

Call a pinning configuration $(\mathcal{C}, \ell)$ *stable* if ℓ remains pinned when the balls in $\mathcal{C}$ are perturbed by any sufficiently small (distinct) motions that keep ℓ fixed. Theorem 5.8 follows from two observations: (i) in any dimension there exists a finite stable pinning configuration (Figure 7 gives an example for $d = 3$), and (ii) in $\mathbb{R}^d$, any stable pinning configuration has size at least $2d-1$.

Theorem 5.8 also narrows the gap on the Hadwiger numbers:

COROLLARY 5.2. *The Hadwiger number of disjoint balls in* $\mathbb{R}^d$ *is* $2d-1$ *or* $2d$.

Since the example of Figure 6(a) can be embedded in $\mathbb{R}^3$, we also have that not every minimal pinning configuration has the same size.

6. Algorithmic aspects. The problem of computing a line transversal to some given collection of sets, if one exists, has been studied in a variety of situations: segments in the plane [15, 30, 62] and higher dimensions [13], convex polygons in the plane [15, 22], polyhedra in three dimensions [13, 49], translates of a convex set in the plane [29]. For families of balls, the best algorithms have complexity $O(n)$ for n disjoint unit disks in the plane [4, 29], $O(n \log n)$ for n intersecting unit disks in the plane [29] and $O(n^{3+\epsilon})$ for n balls in three dimensions [1]; if the dimension is part of the input, deciding if (intersecting) unit balls have a transversal is NP-hard [59]. We complete these results by:

THEOREM 6.1. *A transversal to* n *disjoint balls with bounded radius disparity in* $\mathbb{R}^d$ *can be computed in randomized* $O(n)$ *time.*

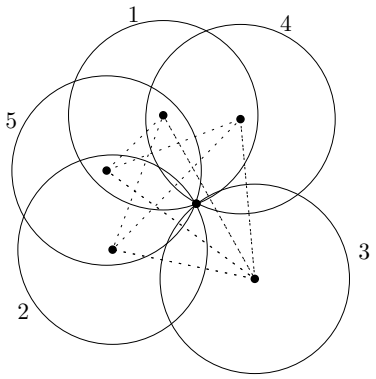


FIG. 7. *Any configuration of* 5 *disjoint balls whose projection along a common tangent is as in the figure is a stable pinning configuration in* $\mathbb{R}^3$.

This is remarkable as there is a $\Omega(n \log n)$ lower bound for this problem for n segments or n unit disks in the plane [12]. The constant in the $O()$ notation depends on the dimension and the radius disparity. The same holds, in any fixed dimension, for thinly distributed collections of balls or more generally any collection for which the number of geometric permutations of any sub-family is bounded.

The next sections briefly recalls the class of *LP-type problems* and uses the connection it bears to Helly-type theorems [4] for deducing Theorem 6.1 from Theorem 5.4.

6.1. Generalized linear programming. The *linear programming* problem, one of the fundamental problems in optimization, consists in maximizing some linear function while satisfying a family of linear equalities and inequalities. Geometrically, it translates into finding a point extremal in some direction (the gradient of the linear function) inside a polytope given as the intersection of half-spaces. Techniques for solving linear programming such as the randomized incremental algorithm of Seidel [69] have been known to solve other problems, for instance computing the smallest enclosing circle of a planar point set. This observation was formalized by Sharir and Welzl [71] in the framework of LP-type problems.

Let $\mathcal{H}$ be a set. Given $F \subset \mathcal{H}$ and $x \in \mathcal{H}$ we denote by $F + x$ and $F - x$ respectively the union and the difference of F and $\{x\}$. An *LP-type problem* is a pair $(\mathcal{H}, w)$ consisting of a set $\mathcal{H}$ and a map $w : 2^{\mathcal{H}} \to \Omega$, where Ω is a totally ordered set with maximal element N, that satisfies for any $F \subset G \subset \mathcal{H}$ and $x \in \mathcal{H}$ the two properties:

- *Monotonicity*: $w(F) \leq w(F + x)$.
- *Locality*: if $w(F) = w(G)$ then

$$w(F + x) \neq w(F) \Leftrightarrow w(G + x) \neq w(G).$$

A subset $F \subset \mathcal{H}$ is a *basis* if it contains no proper subset with the same image under w:

$$\forall x \in F, \quad w(F - x) < w(F).$$

Any set $F \subset \mathcal{H}$ contains a basis B with $w(B) = w(F)$; B is called a *basis of* F. A basis B is *feasible* if $w(B) < N$. The *combinatorial dimension* of an LP-type problem is the maximal cardinality of a feasible basis. Sharir and Welzl [71] showed that if the combinatorial dimension of an LP-type problem $(\mathcal{H}, w)$ is bounded independently of $|\mathcal{H}|$, then a basis of $\mathcal{H}$ can be computed in randomized $O(|\mathcal{H}|)$ time [71].

If $(\mathcal{H}, w)$ is an LP-type problem of combinatorial dimension k, then $w(\mathcal{H})$ is equal to $w(B)$ for some subset $B \subset \mathcal{H}$ of size at most k; thus, for any $\lambda \in \mathbb{R}$, any LP-type problem satisfies the following Helly-type theorem:

$w(\mathcal{H}) \leq \lambda$ if and only if[23] $w(B) \leq \lambda$ for any $B \subset \mathcal{H}$ of size at most k.

This connection goes, in fact, both ways [4] and a large class of Helly-type theorems have a corresponding LP-type problem. The next section applies this correspondence to Theorem 5.4.

6.2. LP-type formulation. Let $\mathcal{C}$ be a collection of disjoint closed balls in $\mathbb{R}^d$. Given a ball X of radius r and a real $\rho \geq 0$ we denote by ρX the ball with same center as X and radius ρr; given a collection F of balls we also denote by ρF the collection $\{\rho X | X \in F\}$. Let $\Omega = [0,1] \cup \{N\}$ where N is maximal and the order on $[0,1]$ is the natural one. The map

$$\phi : \left\{ \begin{array}{ccl} 2^{\mathcal{C}} & \to & \Omega \\ F & \mapsto & \min\left(\left\{\rho \in [0,1] \middle| \rho F \text{ has a transversal}\right\} \bigcup \{N\}\right) \end{array} \right.$$

associates to every sub-collection of balls the amount by which these balls can be "deflated" and still retain some transversal – possibly N if the sub-collection had no transversal to begin with. Theorem 3.1 implies that if $\phi(F) < N$ then $\phi(F)F$ has only finitely many transversals. If $\mathcal{C}$ is not in generic position, there may be more than one such transversal and this implies that $(\mathcal{C}, \phi)$ may violate the locality condition (see Figure 8). Simply put, the system of transversals to balls doesn't meet the "unique minimum property" of Amenta [4]. This can be taken care of as follows[24]. Let $\nu(F)$ denote the number of transversals to $\phi(F)F$, with the convention that $\nu(F) = 0$ whenever $\phi(F) = N$. Define Ω' as $([0,1] \times \mathbb{Z}) \cup \{(N, 0)\}$, ordered lexicographically, and $w = (\phi, \nu) : 2^{\mathcal{C}} \to \Omega'$.

LEMMA 6.1. *$(\mathcal{C}, w)$ is a LP-type problem.*

Proof. Let $F, G \subset \mathcal{C}$ and $x \in \mathcal{C}$. If $\phi(F) < N$ then $w(F + x) = w(F)$ if and only if x intersects every transversal to $\phi(F)F$, so the monotonicity follows. If $F \subset G$ and $w(F) = w(G)$ then $\phi(F)F$ and $\phi(G)G$ have exactly the same set of transversals, and the locality follows. □

[23] The other direction follows from the monotonicity property.

[24] Amenta [4] asserts that a "standart perturbation argument" can also be used.

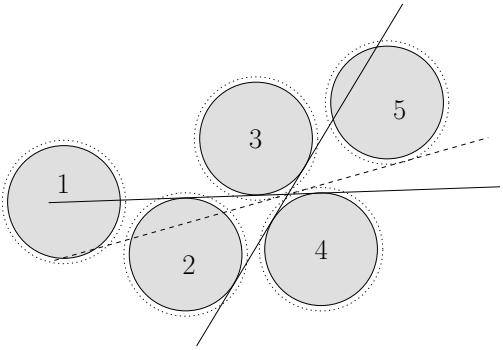


FIG. 8. *ϕ may not fulfill the locality condition (with $F = \{2, 3, 4\}$, $G = F + 5$ and $x = 1$).*

To prove Theorem 6.1, it suffices to observe that the combinatorial dimension of $(\mathcal{C}, w)$ is bounded:

LEMMA 6.2. *For all collections $\mathcal{C}$ of disjoint balls in $\mathbb{R}^d$ with radius disparity at most γ the combinatorial dimension of $(\mathcal{C}, w)$ is $O\left(d^2\gamma^{\log\gamma}\right)$.*

Proof. Let $F \subset \mathcal{C}$ be a basis and denote by $H(d, \gamma)$ the size of the largest family of disjoint balls in $\mathbb{R}^d$ with radius disparity at most γ that has no transversal and is minimal for this property. Theorem 5.4 gives that:

$$H(d, \gamma) = O\left(d^2\gamma^{\log\gamma}\right).$$

Define:

$$\rho = \max\{\phi(B) | B \subset F, B \neq F\}, \quad \text{and}$$

$$\mu = \max\{\nu(B) | B \subset F, B \neq F, \phi(B) = \phi(F)\}.$$

If $\rho \neq \phi(F)$ then for any $\eta \in (\rho, \phi(F))$, the family ηF has no transversal but all its proper subsets do, and thus $|F| \leq H(n, \gamma)$. If $\rho = \phi(F)$ then let B be a basis contained in F such that $\phi(B) = \phi(F)$ and $\nu(B) = \mu$. By definition of μ, for any proper subset B' of B we have $\phi(B') \neq \phi(B)$ and so the previous argument yields that B has size at most $H(d, \gamma)$. Each transversal to $\phi(F)B$ that is not a transversal to $\phi(F)F$ misses some ball $\phi(F)X$ with $X \in F \setminus B$. Thus, since F is a basis, its size is at most $|B| + \mu - \nu(F)$. It follows from Theorem 4.1 that $\mu = O\left(\gamma^{\log\gamma}\right)$ and the statement follows. ∎

REMARK 6.1. The same technique yields that the combinatorial dimension is at most $2d - 1$ for families of balls with a separation set of size 1 (using Theorem 5.2) and $4d - 1$ for families of disjoint unit balls (using Theorem 5.4).

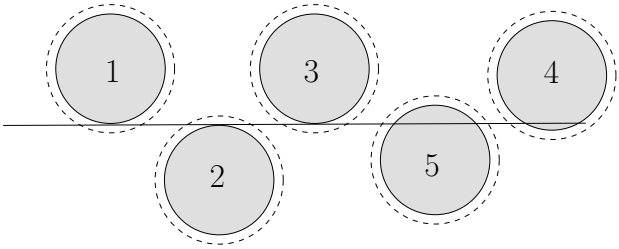


FIG. 9. *For order-respecting transversals the locality condition is not satisfied (with* $F = \{1,2,3\}$, $G = F+4$ *and* $x = 5$*).*

REMARK 6.2. If one defines

$$\phi : \begin{cases} 2^{\mathcal{C}} & \to & \Omega \\ F & \mapsto & \min\left(\left\{\rho \in [0,1] \middle| \rho F \text{ has an order-respecting transversal}\right\} \bigcup \{N\}\right) \end{cases}$$

then the problem $(\mathcal{C}, \phi)$ does not satisfy the locality assumption (see Figure 9). In this case, Theorem 3.2 ensures that the unique minimum property is satisfied.

7. Some open problems. To conclude this overview, we highlight a few of the many questions that remain open.

1. Geometric permutations. What is the asymptotic behavior of the maximum number of geometric permutations of n disjoint convex sets in $\mathbb{R}^d$? The gap between the $\Omega(n^{d-1})$ lower bound [73] and the $O(n^{2d-2})$ upper bound [78] was closed for disjoint balls [73] and fat objects [54] in $\mathbb{R}^d$ and narrowed for sets of bounded description complexity in three dimensions [55]. Also, what is the number of geometric permutations of $n \in \{4, \ldots, 9\}$ disjoint unit balls or to few disjoint balls with bounded radius disparity in dimension $d \geq 3$? A better grasp of these questions may be required to improve the current upper-bounds on the Helly number.

2. Hadwiger number of disjoint balls in $\mathbb{R}^3$. Are the pinning and Hadwiger numbers equal in 3 dimensions? In the plane, the argument used in Theorem 5.1 can be refined to prove that they are; intuitively, case analysis shows that if three objects pin an order-respecting transversal that does not intersect a fourth one, then three of the objects have no order-respecting transversal. Since this analysis exploits the fact that in two dimensions lines are also hyperplanes, it is not clear whether it generalizes to higher dimensions.

3. Pinning number of convex sets. Are there arbitrarily large minimal pinning configurations of convex sets in $\mathbb{R}^d$, or are the corresponding pinning numbers also bounded? Note that there are minimal pinning configurations of size six in $\mathbb{R}^3$ if the objects are not required to be *strictly* convex (see Figure 10).

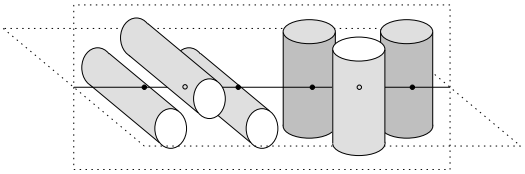


FIG. 10. *A minimal pinning configuration consisting of* 6 *pieces of cylinder: the line is constrained to remain in two planes by triples of cylinders.*

Acknowledgements. The author is grateful to Boris Aronov, Ciprian Borcea, Otfried Cheong, Olivier Devillers, Hazel Everett, Andreas Holmsen and Sylvain Petitjean for helpful discussions.

REFERENCES

[1] P.K. AGARWAL, B. ARONOV, AND M. SHARIR, *Line transversals of balls and smallest enclosing cylinders in three dimensions*, Discrete & Computational Geometry, **21** (1999), pp. 373–388.

[2] N. ALON AND G. KALAI, *Bounding the piercing number*, Discrete & Computational Geometry, **13** (1995), pp. 245–256.

[3] G. AMBRUS, A. BEZDEK, AND F. FODOR, *A Helly-type transversal theorem for n-dimensional unit balls*, Archiv der Mathematik, **86** (2006), pp. 470–480.

[4] N. AMENTA, *Helly-type theorems and generalized linear programming*, Discrete & Computational Geometry, **12** (1994), pp. 241–261.

[5] ———, *A new proof of an interesting Helly-type theorem*, Discrete & Computational Geometry, **15** (1996), pp. 423–427.

[6] B. ARONOV AND S. SMORODINSKY, *On geometric permutations induced by lines transversal through a fixed point*, Discrete & Computational Geometry, **34** (2005), pp. 285–294.

[7] A. ASINOWSKI, *Common transversals and geometric permutations*, master thesis, Technion IIT, Haifa, 1999.

[8] A. ASINOWSKI, A. HOLMSEN, AND M. KATCHALSKI, *The triples of geometric permutations for families of disjoint translates*, Discrete Mathematics, **241** (2001), pp. 23–32.

[9] A. ASINOWSKI, A. HOLMSEN, M. KATCHALSKI, AND H. TVERBERG, *Geometric permutations of large families of translates*, in Discrete and Computational Geometry: The Goodman-Pollack Festschrift, B. Aronov, S. Basu, J. Pach, and M. Sharir, eds., Vol. 25 of Algorithms and Combinatorics, Springer-Verlag, 2003, pp. 157–176.

[10] A. ASINOWSKI AND M. KATCHALSKI, *Forbidden families of geometric permutations in* $\mathbb{R}^d$, Discrete & Computational Geometry, **34** (2005), pp. 1–10.

[11] ———, *The maximal number of geometric permutations for n disjoint translates of a convex set in* $\mathbb{R}^3$ *is* $\omega(n)$, Discrete & Computational Geometry, **35** (2006), pp. 473–480.

[12] D. AVIS, J.-M. ROBERT, AND R. WENGER, *Lower bounds for line stabbing*, Information processing letters, **33** (1989), pp. 59–62.

[13] D. AVIS AND R. WENGER, *Polyhedral line transversals in space*, Discrete & Computational Geometry, **3** (1988), pp. 257–265.

[14] M. BERN AND D. EPPSTEIN, *Multivariate regression depth*, Discrete & Computational Geometry, **28** (2002), pp. 1–17.

[15] B. Bhattacharya, J. Czyzowicz, P. Egyed, G. Toussaint, I. Stojmenovic, and J. Urrutia, *Computing shortest transversals of sets*, International Journal of Computational Geometry and Applications, **2** (1992), pp. 417–435.

[16] C. Borcea, X. Goaoc, and S. Petitjean, *Line transversals to disjoint balls*, Discrete & Computational Geometry, **1–3** (2008), pp. 158–173.

[17] O. Cheong, X. Goaoc, and A. Holmsen, *Hadwiger and Helly-type theorems for disjoint unit spheres in* $\mathbb{R}^3$, in Proc. 20th Ann. Symp. on Computational Geometry, 2005, pp. 10–15.

[18] ———, *Lower bounds for pinning lines by balls.* Manuscript, 2008.

[19] O. Cheong, X. Goaoc, A. Holmsen, and S. Petitjean, *Hadwiger and Helly-type theorems for disjoint unit spheres*, Discrete & Computational Geometry, **1–3** (2008), pp. 194–212.

[20] O. Cheong, X. Goaoc, and H.-S. Na, *Disjoint unit spheres admit at most two line transversals*, in Proc. 11th Annu. European Sympos. Algorithms, Vol. 2832 of Lecture Notes in Computer Science, 2003, pp. 127–135.

[21] ———, *Geometric permutations of disjoint unit spheres*, Computational Geometry: Theory & Applications, **30** (2005), pp. 253–270.

[22] F.Y.L. Chin and F.L. Wang, *Efficient algorithm for transversal of disjoint convex polygons*, Information processing letters, **83** (2002), pp. 141–144.

[23] L. Danzer, *Über ein Problem aus der kombinatorischen Geometrie*, Archiv der Mathematik, (1957).

[24] L. Danzer, B. Grünbaum, and V. Klee, *Helly's theorem and its relatives*, in Convexity, V. Klee, ed., Proc. of Symposia in Pure Math., Amer. Math. Soc., 1963, pp. 101–180.

[25] H. Debrunner, *Helly type theorems derived from basic singular homology*, American Mathematical Monthly, **77** (1970), pp. 375–380.

[26] F. Durand, *A multidisciplinary survey of visibility*, in ACM SIGGRAPH Course Notes: Visibility, Problems, Techniques, and Applications, 2000.

[27] J. Eckhoff, *Helly, Radon and Caratheodory type theorems*, in Handbook of Convex Geometry, J.E. Goodman and J. O'Rourke, eds., North Holland, 1993, pp. 389–448.

[28] H. Edelsbrunner and M. Sharir, *The maximum number of ways to stab n convex non-intersecting sets in the plane is* $2n-2$, Discrete & Computational Geometry, **5** (1990), pp. 35–42.

[29] P. Egyed and R. Wenger, *Stabbing pairwise disjoint translates in linear time*, in Proc. 5th Symposium on Computational Geometry, 1989, pp. 364–369.

[30] H. Everett, J.-M. Robert, and M. van Kreveld, *An optimal algorithm for the* $(\leq k)$*-levels, with applications to separation and transversal problems*, in Proc. 9th Symposium on Computational Geometry, 1993, pp. 38–46.

[31] J.E. Goodman, *When is a set of lines in space convex?*, Notices of the AMS, **45** (1998), pp. 222–232.

[32] J.E. Goodman, A. Holmsen, R. Pollack, K. Ranestad, and F. Sottile, *Cremona convexity, frame convexity, and a theorem of Santaló*, Advances in Geometry, **6** (2006), pp. 301–322.

[33] J.E. Goodman and R. Pollack, *Foundations of a theory of convexity on affine grassmann manifolds*, Mathematika, **42** (1995), pp. 308–328.

[34] ———, *Geometric transversal theory*, in Encyclopaedia of Mathematics, Springer-Verlag, Heidelberg, Germany, 2002. `http://eom.springer.de/g/g130050.htm`.

[35] J.E. Goodman, R. Pollack, and R. Wenger, *Geometric transversal theory*, in New Trends in Discrete and Computational Geometry, J. Pach, ed., Vol. 10 of Algorithms and Combinatorics, Springer-Verlag, Heidelberg, Germany, 1993, pp. 163–198.

[36] E. Greenstein and M. Sharir, *The space of line transversals to pairwise disjoint balls in* $\mathbb{R}^3$. manuscript, 2005.

[37] B. Grünbaum, *On common transversals*, Archiv der Mathematik, 9 (1958), pp. 465–469.

[38] ———, *Common transversals for families of sets*, Journal of the London Mathematical Society, **35** (1960), pp. 408–416.
[39] B. Grünbaum and T. Motzkin, *On components in some families of sets*, Proceedings of the American Mathematical Society, **12** (1961), pp. 607–613.
[40] H. Hadwiger, *Problem 107. Nieuw Archiv Wiskunde*, (3)4:57, 1956; Solution. *Wiskundige Opgaven*, **20** (1957), pp. 27–29.
[41] ———, *Über eibereiche mit gemeinsamer treffgeraden*, Portugal Math., **6** (1957), pp. 23–29.
[42] E. Helly, *Über Mengen konvexer Körper mit gemeinschaftlichen Punkten*, Jahresbericht Deutsch. Math. Verein., **32** (1923), pp. 175–176.
[43] ———, *Über Systeme von abgeschlossenen Mengen mit gemeinschaftlichen Punkten*, Monaths. Math. und Physik, **37** (1930), pp. 281–302.
[44] A. Holmsen, *Recent progress on line transversals to families of translated ovals*, in Computational Geometry - Twenty Years Later, J.E. Goodman, J. Pach, and R. Pollack, eds., AMS, 2008, pp. 283–298.
[45] A. Holmsen, M. Katchalski, and T. Lewis, *A Helly-type theorem for line transversals to disjoint unit balls*, Discrete & Computational Geometry, **29** (2003), pp. 595–602.
[46] A. Holmsen and J. Matoušek, *No Helly theorem for stabbing translates by lines in* $\mathbb{R}^d$, Discrete & Computational Geometry, **31** (2004), pp. 405–410.
[47] Y. Huang, J. Xu, and D.Z. Chen, *Geometric permutations of high dimensional spheres*, in Proc. 12th ACM-SIAM Sympos. Discrete Algorithms, 2001, pp. 244–245.
[48] ———, *Geometric permutations of high dimensional spheres*, Computational Geometry: Theory & Applications, **29** (2004), pp. 47–60.
[49] J. Jaromczyk and M. Kowaluk, *Skewed projections with an application to line stabbing in* $\mathbb{R}^3$, in Proc. 4th Conference on Computational Geometry, 1988, pp. 362–370.
[50] M. Katchalski, T. Lewis, and A. Liu, *Geometric permutations and common transversals*, Discrete & Computational Geometry, **1** (1986), pp. 371–377.
[51] ———, *Geometric permutations of disjoint translates of convex sets*, Discrete Mathematics, 65 (1987), pp. 249–259.
[52] ———, *The different ways of stabbing disjoint convex sets*, Discrete & Computational Geometry, **7** (1992), pp. 197–206.
[53] M. Katchalski, S. Suri, and Y. Zhou, *A constant bound for geometric permutations of disjoint unit balls*, Discrete & Computational Geometry, **29** (2003), pp. 161–173.
[54] M.J. Katz and K.R. Varadarajan, *A tight bound on the number of geometric permutations of convex fat objects in* $\mathbb{R}^d$, Discrete & Computational Geometry, **26** (2001), pp. 543–548.
[55] V. Koltun and M. Sharir, *The partition technique for overlays of enveloppes*, SIAM Journal of Computing, **32** (2003), pp. 841–863.
[56] *The Maple System*. Waterloo Maple Software. `http://www.maplesoft.com`.
[57] J. Matoušek, *Lectures on Discrete Geometry*, Springer-Verlag, 2002.
[58] J. Matousēk, *A Helly-type theorem for unions of convex sets*, Discrete & Computational Geometry, **18** (1997), pp. 1–12.
[59] N. Megiddo, *On the complexity of some geometric problems in unbounded dimension*, Journal of Symbolic Computation, **10** (1990), pp. 327–334.
[60] G. Megyesi and F. Sottile, *The envelope of lines meeting a fixed line and tangent to two spheres*, Discrete & Computational Geometry, **33** (2005), pp. 617–644. arXiv math.AG/0304346.
[61] J. Mitchell and M. Sharir, *New results on shortest paths in three dimensions*, in Proc. 20th Symposium on Computational Geometry, 2004, pp. 124–133.
[62] J. O'Rourke, *An on-line algorithm for fitting straight lines between data ranges*, Communications of the ACM, **24** (1981), pp. 574–579.

[63] J. Pach and M. Sharir, *Combinatorial Geometry with Algorithmic Applications – The Alcala Lectures*. Alcala (Spain), August 31 - September 5, 2006.

[64] M. Pellegrini, *Ray shooting and lines in space*, in Handbook of Discrete & Computational Geometry, J. E. Goodman and J. O'Rourke, eds., CRC Press LLC, 2004, ch. 37, pp. 839–856.

[65] H. Pottmann and J. Wallner, *Computational Line Geometry*, Springer, 2001.

[66] C.V. Robinson, *Spherical theorems of Helly type and congruence indices of spherical caps*, American Journal of Mathematics, **64** (1942), pp. 260–272.

[67] P. Rousseeuw and M. Hubert, *Regression depth*, J. Amer. Stat. Assoc., **94** (1999), pp. 388–402.

[68] L. Santaló, *Un theorema sobre conjuntos de paralelepipedos de aristas paralelas*, Publ. Inst. Mat. Univ. Nat. Litoral, **2** (1940), pp. 49–60.

[69] R. Seidel, *Small-dimensional linear programming and convex hulls made easy*, Discrete & Computational Geometry, **6** (1991), pp. 423–434.

[70] M. Sharir and S. Smorodinsky, *On neighbors in geometric permutations*, Discrete Mathematics, **268** (2003), pp. 327–335.

[71] M. Sharir and E. Welzl, *A combinatorial bound for linear programming and related problems*, in Proc. 9th Sympos. on Theo. Aspects of Comp. Science, 1992, pp. 569–579.

[72] S. Smorodinsky, *Geometric permutations and common transversals*, master thesis, Tel Aviv University, Haifa, 1998.

[73] S. Smorodinsky, J.S.B. Mitchell, and M. Sharir, *Sharp bounds on geometric permutations for pairwise disjoint balls in* $\mathbb{R}^d$, Discrete & Computational Geometry, **23** (2000), pp. 247–259.

[74] F. Sottile and T. Theobald, *Line problems in nonlinear computational geometry*, in Computational Geometry - Twenty Years Later, J.E. Goodman, J. Pach, and R. Pollack, eds., AMS, 2008, pp. 411–432.

[75] H. Tverberg, *Proof of Grünbaum's conjecture on common transversals for translates*, Discrete & Computational Geometry, **4** (1989), pp. 191–203.

[76] P. Vincensini, *Figures convexes et variétés linéaires de l'espace euclidien à n dimensions*, Bull. Sci. Math., **59** (1935), pp. 163–174.

[77] R. Wenger, *A generalization of hadwiger's transversal theorem to intersecting sets*, Discrete & Computational Geometry, **5** (1990), pp. 383–388.

[78] ———, *Upper bounds on geometric permutations for convex sets*, Discrete & Computational Geometry, **5** (1990), pp. 27–33.

[79] ———, *Helly-type theorems and geometric transversals*, in Handbook of Discrete & Computational Geometry, J. E. Goodman and J. O'Rourke, eds., CRC Press LLC, Boca Raton, FL, 2nd ed., 2004, ch. 4, pp. 73–96.

[80] Y. Zhou and S. Suri, *Shape sensitive geometric permutations*, in Proc. 12th ACM-SIAM Sympos. Discrete Algorithms, 2001, pp. 234–243.

[81] ———, *Geometric permutations of balls with bounded size disparity*, Computational Geometry: Theory & Applications, **26** (2003), pp. 3–20.

ALGEBRAIC GEOMETRY AND KINEMATICS

MANFRED L. HUSTY* AND HANS-PETER SCHRÖCKER*†

Abstract. In this overview paper we show how problems in computational kinematics can be translated into the language of algebraic geometry and subsequently solved using techniques developed in this field. The idea to transform kinematic features into the language of algebraic geometry is old and goes back to Study. Recent advances in algebraic geometry and symbolic computation gave the motivation to resume these ideas and make them successful in the solution of kinematic problems. It is not the aim of the paper to provide detailed solutions, but basic accounts to the used tools and examples where these techniques were applied within the last years. We start with Study's kinematic mapping and show how kinematic entities can be transformed into algebraic varieties. The transformations in the image space that preserve the kinematic features are introduced. The main topic are the definition of constraint varieties and their application to the solution of direct and inverse kinematics of serial and parallel robots. We provide a definition of the degree of freedom of a mechanical system that takes into account the geometry of the device and discuss singularities and global pathological behavior of selected mechanisms. In a short paragraph we show how the developed methods are applied to the synthesis of mechanical devices.

Key words. Computational kinematics, kinematic mapping, constraint variety, serial robot, parallel robot, singularity, self-motion, synthesis of mechanisms.

AMS(MOS) subject classifications. 70B15, 53A17.

1. Introduction. This chapter is devoted to the application of algebraic geometry in computational kinematics. We study the motions of mechanical devices (linkages or robots) and their relation to individual joint parameters, we design mechanisms such that they can perform certain prescribed tasks and we explain "pathological" or surprising behavior of mechanisms by relating them to well-known concepts of algebraic geometry.

The theoretical foundations of our topic are old and date back to the 19th century and beyond. It was, however, only the power of modern computers, in particular algebraic manipulation systems and advances in numerical computation [25], that allowed the application to "real" problems in computational kinematics and mechanism science. Many previously hopeless problems can nowadays be solved in fractions of a second. On the other hand, scientists and engineers are constantly attacking new problems just at the edge of feasibility.

2. Kinematic mapping. A fundamental concept of relating mechanical structures with algebraic varieties is Study's kinematic mapping [26, 27]. It associates to every Euclidean displacement γ a point $\mathbf{c}$ in real projective space P^7 of dimension seven or, more precisely, a point on the Study quadric

* University Innsbruck, Institute of Basic Sciences in Engineering, Unit Geometry and CAD, Technikerstraße 13, A6020 Innsbruck, Austria, http://geometrie.uibk.ac.at (manfred.husty@uibk.ac.at).

† hans-peter.schroecker@uibk.ac.at.

I.Z. Emiris et al. (eds.), *Nonlinear Computational Geometry*, The IMA Volumes in Mathematics and its Applications 151, DOI 10.1007/978-1-4419-0999-2_4,

$S \subset P^7$. Sometimes it will be necessary to use also its complex extension $P^7(\mathbb{C})$. There exist other kinematic mappings besides Study's ([3, 9, 21, 30]) but these topics are beyond the scope of the present text.

Within the kinematics community more often four by four matrices incorporating translational and rotational part of the motion are used (2.2). Matrix elements are the design parameters of the mechanism (often called Denavit-Hartenberg parameters) and sines and cosines of the motion parameters. To move to algebra one either uses tangent half substitution transforming sines and cosines to algebraic values or adds the identity $\sin^2 + \cos^2 = 1$.

A formal definition of Study's kinematic mapping is given below in Subsection 2.1. Our description is based on the original works of Study [26, 27]. These are comprehensive and deep but not always easily readable texts and, unfortunately, only available in German. Modern references on the same topic include [15] or [23].

2.1. Study's kinematic mapping. Euclidean three space is the three dimensional real vector space $\mathbb{R}^3$ together with the usual scalar product $\mathbf{x}^T\mathbf{y} = \sum_{i=1}^{3} x_i y_i$. A Euclidean displacement is a mapping

$$\gamma\colon \mathbb{R}^3 \to \mathbb{R}^3, \quad \mathbf{x} \mapsto \mathbf{A}\mathbf{x} + \mathbf{a} \tag{2.1}$$

where $\mathbf{A} \in \mathrm{SO}(3)$ is a proper orthogonal three by three matrix and $\mathbf{a} \in \mathbb{R}^3$ is a vector. The entries of $\mathbf{A}$ fulfill the well-known orthogonality condition $\mathbf{A}^T \cdot \mathbf{A} = \mathbf{I}_3$, where $\mathbf{I}_3$ is the three by three identity matrix.

The group of all Euclidean displacements is denoted by SE(3). It is a convenient convention to write Equation (2.1) as product of a four by four matrix and a four dimensional vector according to

$$\begin{bmatrix} 1 \\ \mathbf{x} \end{bmatrix} \mapsto \begin{bmatrix} 1 & \mathbf{o}^T \\ \mathbf{a} & \mathbf{A} \end{bmatrix} \cdot \begin{bmatrix} 1 \\ \mathbf{x} \end{bmatrix}. \tag{2.2}$$

Study's kinematic mapping $\varkappa$ maps an element α of SE(3) to a point $\mathbf{x} \in P^7$. If the homogeneous coordinate vector of $\mathbf{x}$ is $[x_0 : x_1 : x_2 : x_3 : y_0 : y_1 : y_2 : y_3]^T$, the kinematic pre-image of $\mathbf{x}$ is the displacement α described by the transformation matrix

$$\frac{1}{\Delta}\begin{bmatrix} \Delta & 0 & 0 & 0 \\ p & x_0^2 + x_1^2 - x_2^2 - x_3^2 & 2(x_1x_2 - x_0x_3) & 2(x_1x_3 + x_0x_2) \\ q & 2(x_1x_2 + x_0x_3) & x_0^2 - x_1^2 + x_2^2 - x_3^2 & 2(x_2x_3 - x_0x_1) \\ r & 2(x_1x_3 - x_0x_2) & 2(x_2x_3 + x_0x_1) & x_0^2 - x_1^2 - x_2^2 + x_3^2 \end{bmatrix} \tag{2.3}$$

where

$$\begin{aligned} p &= 2(-x_0y_1 + x_1y_0 - x_2y_3 + x_3y_2), \\ q &= 2(-x_0y_2 + x_1y_3 + x_2y_0 - x_3y_1), \\ r &= 2(-x_0y_3 - x_1y_2 + x_2y_1 + x_3y_0), \end{aligned} \tag{2.4}$$

and $\Delta = x_0^2 + x_1^2 + x_2^2 + x_3^2$. The lower three by three sub-matrix is a proper orthogonal matrix if and only if

$$x_0y_0 + x_1y_1 + x_2y_2 + x_3y_3 = 0 \tag{2.5}$$

and not all x_i are zero. When these conditions are fulfilled we call $[x_0 : \ldots : y_3]^T$ the *Study parameters* of the displacement α.

The important relation (2.5) defines a quadric $S \subset P^7$ and the range of $\varkappa$ is this quadric minus the three dimensional subspace defined by

$$E\colon x_0 = x_1 = x_2 = x_3 = 0. \tag{2.6}$$

We call S the *Study quadric* and E the *exceptional* or *absolute generator*.

The parameterization (2.3) of SE(3) may look rather artificial and complicated but it has an important feature: *The composition of displacements in Study parameters is bilinear* (see Subsection 2.2). In [26] Study shows that

- this requirement cannot be fulfilled with a smaller number of parameters and
- the representation of Euclidean displacements is unique, up to linear parameter transformations and transformations via identically fulfilled relations between the parameters.

Moreover, the Study parameters are closely related to the ring of *biquaternions* or *dual quaternions* as we shall rather say. This will be on the agenda in the coming Subsection 2.4. But first, we fill a gap that is so far missing in our exposée.

For the description of a mechanical device in P^7 we usually need the inverse of the map given by Equations (2.3) and (2.4), that is, we need to know how to compute the Study parameters from the entries of the matrix $\mathbf{A} = [a_{ij}]_{i,j=1,\ldots,3}$ and the vector $\mathbf{a} = [a_1, a_2, a_3]^T$. Mostly in kinematics literature a rather complicated and not singularity-free procedure, based on the Cayley transform of a skew symmetric matrix into an orthogonal matrix (see [8]), is used. The best way of doing this was, however, already known to Study himself. He showed that the homogeneous quadruple $x_0 : x_1 : x_2 : x_3$ can be obtained from at least one of the following proportions:

$$\begin{aligned} x_0 : x_1 : x_2 : x_3 &= 1+a_{11}+a_{22}+a_{33} : a_{32}-a_{23} : a_{13}-a_{31} : a_{21}-a_{12} \\ &= a_{32}-a_{23} : 1+a_{11}-a_{22}-a_{33} : a_{12}+a_{21} : a_{31}+a_{13} \\ &= a_{13}-a_{31} : a_{12}+a_{21} : 1-a_{11}+a_{22}-a_{33} : a_{23}+a_{32} \\ &= a_{21}-a_{12} : a_{31}+a_{13} : a_{23}-a_{32} : 1-a_{11}-a_{22}+a_{33}. \end{aligned} \tag{2.7}$$

In general, all four proportions of (2.7) yield the same result. If, however, $1 + a_{11} + a_{22} + a_{33} = 0$ the first proportion yields $0 : 0 : 0 : 0$ and is invalid. We can use the second proportion instead as long as $a_{22} + a_{33}$ is different from zero. If this happens we can use the third proportion unless $a_{11} + a_{33} = 0$. In this last case we resort to the last proportion which yields

$0:0:0:1$. Having computed the first four Study parameters the remaining four parameters $y_0 : y_1 : y_2 : y_3$ can be computed from

$$\begin{aligned} 2y_0 &= a_1x_1 + a_2x_2 + a_3x_3, \\ 2y_1 &= -a_1x_0 + a_3x_2 - a_2x_3, \\ 2y_2 &= -a_2x_0 - a_3x_1 + a_1x_3, \\ 2y_3 &= -a_3x_0 + a_2x_1 - a_1x_2. \end{aligned} \tag{2.8}$$

EXAMPLE 1. *A rotation about the z-axis through the angle φ is described by the matrix*

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\varphi & -\sin\varphi & 0 \\ 0 & \sin\varphi & \cos\varphi & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{2.9}$$

Its kinematic image, computed via (2.7) *and* (2.8) *is*

$$\mathbf{r} = [1 + \cos\varphi : 0 : 0 : \sin\varphi : 0 : 0 : 0 : 0]^T. \tag{2.10}$$

As φ varies in $[0, 2\pi)$, $\mathbf{r}$ describes a straight line on the Study quadric which reads after algebraization

$$\mathbf{r} = [1 : 0 : 0 : u : 0 : 0 : 0 : 0]^T. \tag{2.11}$$

EXERCISE 2.1. *Compute the Study representation of the translations in direction of the z-axis.*

2.2. Fixed and moving frame. Suppose that $\alpha\colon \mathbf{x} \mapsto \mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{a}$ is a Euclidean displacement. The vectors $\mathbf{x}$ and $\mathbf{y}$ are elements of $\mathbb{R}^3$ but in kinematics it is advantageous to consider them as elements of two distinct copies of $\mathbb{R}^3$, called the *moving space* and the *fixed space*. The description of α in Study parameters depends on the choice of coordinate frames – *moving frame* and *fixed* or *base frame* – in both spaces. In kinematics, the moving frame is the space attached to a mechanism's output link, and the fixed space is the space where the mechanism itself is positioned (see Subsection 3).

Both types of transformations induce transformations of the Study quadric and thus impose a geometric structure on P^7. Kinematic mapping is constructed such that these transformations act linearly on the Study parameters (that is, they are projective transformations in P^7). We are going to compute their coordinate representations.

Consider a Euclidean displacement described by a four by four transformation matrix $\mathbf{X}$, as in (2.3). It maps a point $(1, \mathbf{a})^T$ to $(1, \mathbf{a}') = \mathbf{X} \cdot (1, \mathbf{a})^T$. Now we change coordinate frames in fixed and moving space and compute the matrix $\mathbf{Y}$ such that $(1, \mathbf{b}')^T = \mathbf{Y} \cdot (1, \mathbf{b})^T$ is the representation of the

displacement in the new fixed coordinate frame and the *old* moving coordinate frame. This is slightly different from the typical change of coordinates known from linear algebra where one describes the new transformation in terms of new coordinates in *both* spaces but more suitable for application in kinematics, in particular for describing the position of the end effector tool or for concatenation of simple mechanisms. If the changes of coordinates in fixed and moving frame are described by

$$(1,\mathbf{a})^T = \mathbf{M}\cdot(1,\mathbf{b})^T, \quad (1,\mathbf{b}')^T = \mathbf{F}\cdot(1,\mathbf{a}')^T, \tag{2.12}$$

we have $\mathbf{Y} = \mathbf{F}\cdot\mathbf{X}\cdot\mathbf{M}$. Denote now by $\mathbf{y}$, $\mathbf{x}$, $\mathbf{f} = [f_0,\ldots,f_7]^T$ and $\mathbf{m} = [m_0,\ldots,m_7]^T$ the corresponding Study vectors. Straightforward computation (or skillful use of dual quaternions, see Subsection 2.4) yields

$$\mathbf{y} = \mathbf{T}_f\mathbf{T}_m\mathbf{x}, \quad \mathbf{T}_m = \begin{bmatrix}\mathbf{A} & \mathbf{O}\\ \mathbf{B} & \mathbf{A}\end{bmatrix}, \quad \mathbf{T}_f = \begin{bmatrix}\mathbf{C} & \mathbf{O}\\ \mathbf{D} & \mathbf{C}\end{bmatrix}, \tag{2.13}$$

where

$$\begin{aligned}
\mathbf{A} &= \begin{bmatrix} m_0 & -m_1 & -m_2 & -m_3\\ m_1 & m_0 & m_3 & -m_2\\ m_2 & -m_3 & m_0 & m_1\\ m_3 & m_2 & -m_1 & m_0\end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} m_4 & -m_5 & -m_6 & -m_7\\ m_5 & m_4 & m_7 & -m_6\\ m_6 & -m_7 & m_4 & m_5\\ m_7 & m_6 & -m_5 & m_4\end{bmatrix}\\
\mathbf{C} &= \begin{bmatrix} f_0 & -f_1 & -f_2 & -f_3\\ f_1 & f_0 & -f_3 & f_2\\ f_2 & f_3 & f_0 & -f_1\\ f_3 & -f_2 & f_1 & f_0\end{bmatrix}, \quad \mathbf{D} = \begin{bmatrix} f_4 & -f_5 & -f_6 & -f_7\\ f_5 & f_4 & -f_7 & f_6\\ f_6 & f_7 & f_4 & -f_5\\ f_7 & -f_6 & f_5 & f_4\end{bmatrix},
\end{aligned} \tag{2.14}$$

and $\mathbf{O}$ is the four by four zero matrix.

The matrices $\mathbf{T}_m$ and $\mathbf{T}_f$ commute and they induce transformations of P^7 that leave fixed the Study quadric S, the exceptional generator E, and the *exceptional* or *absolute quadric* $F \subset E$, defined by the equations

$$F : x_0 = x_1 = x_2 = x_3 = 0, \quad y_0^2 + y_1^2 + y_2^2 + y_3^2 = 0. \tag{2.15}$$

The quadrics S and F and the three space E are special objects in the geometry of the kinematic image space. Later we will describe a mechanism by a subvariety V of P^7. A non-generic position of V with respect to these objects distinguishes its kinematic properties from a projectively equivalent subvariety W.

Example 2. *In Example 1 we saw that the kinematic image of a continuous rotation about the z-axis is a straight line on the Study quadric. From the considerations in this section it follows that the kinematic image of a continuous rotation about an arbitrary axis is a straight line.*

Consider now a straight line $L \subset \mathbb{R}^3$ and denote the foot points of the common perpendicular to L and the z-axis Z by $\mathbf{f}_L$ and $\mathbf{f}_Z$. Assuming the common perpendicular of Z and L points in direction of the x-axis, the

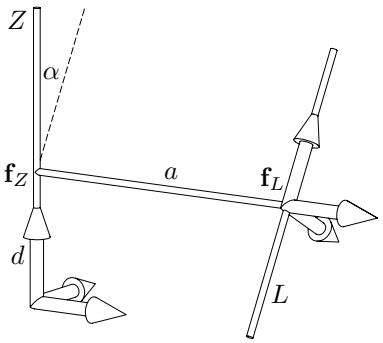


FIG. 1. *Relative position of two lines Z and L (courtesy Martin Pfurner).*

relative position of L with respect to Z can be specified by the z-coordinate d of $\mathbf{f}_Z$*, the distance a between* $\mathbf{f}_Z$ *and* $\mathbf{f}_L$ *and the angle* α *between Z and L (Figure 1). The numbers d, a and* α *are called the Denavit-Hartenberg parameters of the relative position of the line L with respect to the z-axis (see [28, Section 2.3]).*

The displacement that transforms the standard coordinate frame to the coordinate frame with origin $\mathbf{f}_L$*, x-axis in direction of the common perpendicular, and z-axis in direction of L in matrix form reads*

$$\mathbf{G} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ a & 1 & 0 & 0 \\ 0 & 0 & \cos\alpha & -\sin\alpha \\ d & 0 & \sin\alpha & \cos\alpha \end{bmatrix}. \tag{2.16}$$

Its Study vector is

$$\mathbf{g} = [2\gamma, 2\sin\alpha, 0, 0, a\sin\alpha, -a\gamma, -d\sin\alpha, -d\gamma]^T \tag{2.17}$$

where $\gamma = 1 + \cos\alpha$*. The kinematic image of the rotation about L is* $\mathbf{T}_f \cdot \mathbf{r}$ *where* $\mathbf{T}_f$ *is obtained by substituting the components of* $\mathbf{g}$ *into* (2.13).

EXERCISE 2.2. *Compute the Study representation of all translations in a fixed direction different from* $[0, 0, 1]^T$*.*

2.3. Planar and spherical kinematic mapping. The restriction of Study's kinematic mapping to certain three-spaces on the Study quadric yields elements of two important subgroups of SE(3), the group of planar Euclidean displacements SE(2) and the special orthogonal group SO(3) whose elements are pure rotations without any translational component. Both groups are of relevance in kinematics. Their kinematic mappings will be introduced in this subsection.

The planar Euclidean motion group SE(2) can be embedded into SE(3) by substituting $x_1 = x_2 = y_0 = y_3 = 0$ into (2.3). This yields the matrix parameterization

$$\frac{1}{x_0^2+x_3^2}\begin{bmatrix} x_0^2+x_3^2 & 0 & 0 \\ -2(x_0y_1-x_3y_2) & x_0^2-x_3^2 & -2x_0x_3 \\ -2(x_0y_2+x_3y_1) & 2x_0x_3 & x_0^2-x_3^2 \end{bmatrix} \tag{2.18}$$

of SE(2) (we omit the last row and the last column). The group SE(2) can be considered as kinematic pre-image of the three space $x_1 = x_2 = y_0 = y_3 = 0$, minus its intersection with the exceptional generator E, and we identify this three space with P^3. We describe its points by homogeneous coordinates $[x_0 : x_3 : y_1 : y_2]^T$.

The geometry of P^3 as range of planar kinematic mapping is governed by a change of coordinates in the moving or fixed frame or, equivalently, by its absolute figure consisting of the line $x_0 = x_3 = 0$ (the intersection of P^3 with the exceptional generator E) and the absolute points $[0 : 0 : 1 : \pm\mathrm{i}]^T$ (the intersection of P^3 with the absolute quadric F). This geometry is called *quasielliptic* (see for example [4, p. 399]).

The spherical motion group SO(3) can be embedded into SE(3) via

$$\frac{1}{\Delta}\begin{bmatrix} x_0^2+x_1^2-x_2^2-x_3^2 & 2(x_1x_2-x_0x_3) & 2(x_1x_3+x_0x_2) \\ 2(x_1x_2+x_0x_3) & x_0^2-x_1^2+x_2^2-x_3^2 & 2(x_2x_3-x_0x_1) \\ 2(x_1x_3-x_0x_2) & 2(x_2x_3+x_0x_1) & x_0^2-x_1^2-x_2^2+x_3^2 \end{bmatrix} \tag{2.19}$$

where $\Delta = x_0^2 + x_1^2 + x_2^2 + x_3^2$. It is the kinematic pre-image of the three space $y_0 = y_1 = y_2 = y_3 = 0$. The absolute figure is the exceptional quadric $x_0^2 + x_1^2 + x_2^2 + x_3^2 = 0$ and the corresponding geometry is *elliptic* (see for example [6, Chapter VII]).

2.4. Euclidean displacements and dual quaternions. Kinematic mapping is closely related to quaternion algebra. This relation shall be illustrated in this section. The set of quaternions $\mathbb{H}$ is the vector space $\mathbb{R}^4$ together with the quaternion multiplication

$$\begin{aligned}(a_0,a_1,a_2,a_3)\star(b_0,b_1,b_2,b_3) = (&a_0b_0-a_1b_1-a_2b_2-a_3b_3,\\ &a_0b_1+a_1b_0+a_2b_3-a_3b_2,\\ &a_0b_2-a_1b_3+a_2b_0-a_3b_1,\\ &a_0b_3-a_1b_2-a_2b_1+a_3b_0).\end{aligned} \tag{2.20}$$

The triple $(\mathbb{H}, +, \star)$ (with component wise addition) forms a skew field. The real numbers can be embedded into this field via $x \mapsto (x, 0, 0, 0)$, and vectors $\mathbf{x} \in \mathbb{R}^3$ are identified with quaternions of the shape $(0, \mathbf{x})$.

Every quaternion is a unique linear combination of the four basis quaternions $\mathbf{1} = (1, 0, 0, 0)$, $\mathbf{i} = (0, 1, 0, 0)$, $\mathbf{j} = (0, 0, 1, 0)$, and $\mathbf{k} = (0, 0, 0, 1)$. Their multiplication table is

$\star$	$\mathbf{1}$	$\mathbf{i}$	$\mathbf{j}$	$\mathbf{k}$
$\mathbf{1}$	$\mathbf{1}$	$\mathbf{i}$	$\mathbf{j}$	$\mathbf{k}$
$\mathbf{i}$	$\mathbf{i}$	$-\mathbf{1}$	$\mathbf{k}$	$-\mathbf{j}$
$\mathbf{j}$	$\mathbf{j}$	$-\mathbf{k}$	$-\mathbf{1}$	$\mathbf{i}$
$\mathbf{k}$	$\mathbf{k}$	$\mathbf{j}$	$-\mathbf{i}$	$-\mathbf{1}$

Conjugate quaternion and *norm* are defined as

$$\overline{A} = (a_0, -a_1, -a_2, -a_3), \quad \|A\| = \sqrt{A \star \overline{A}} = \sqrt{a_0^2 + a_1^2 + a_2^2 + a_3^2}. \tag{2.21}$$

Quaternions are closely related to spherical kinematic mapping. Consider a vector $\mathbf{a} = [a_1, a_2, a_3]^T$ and a matrix $\mathbf{X}$ of the shape (2.19). Then the product $\mathbf{b} = \mathbf{X} \cdot \mathbf{a}$ can also be written as

$$\mathbf{b} = \mathbf{a} \star X \star \overline{\mathbf{a}} \tag{2.22}$$

where $X = (x_0, x_1, x_2, x_3)$ and $\|X\| = 1$. In other words, spherical displacements can also be described by unit quaternions and spherical kinematic mapping maps a spherical displacement to the corresponding unit quaternion.

In order to describe general Euclidean displacements we have to extend the concept of quaternions. A *dual quaternion* Q is a quaternion over the ring of dual numbers, that is, it can be written as

$$Q = Q_0 + \varepsilon Q_1, \tag{2.23}$$

where $\varepsilon^2 = 0$. The algebra of dual quaternions has eight basis elements $\mathbf{1}$, $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$, ε, $\varepsilon\mathbf{i}$, $\varepsilon\mathbf{j}$, and $\varepsilon\mathbf{k}$ and the multiplication table

$\star$	$\mathbf{1}$	$\mathbf{i}$	$\mathbf{j}$	$\mathbf{k}$	ε	$\varepsilon\mathbf{i}$	$\varepsilon\mathbf{j}$	$\varepsilon\mathbf{k}$
$\mathbf{1}$	$\mathbf{1}$	$\mathbf{i}$	$\mathbf{j}$	$\mathbf{k}$	ε	$\varepsilon\mathbf{i}$	$\varepsilon\mathbf{j}$	$\varepsilon\mathbf{k}$
$\mathbf{i}$	$\mathbf{i}$	$-\mathbf{1}$	$\mathbf{k}$	$-\mathbf{j}$	$\varepsilon\mathbf{i}$	$-\varepsilon\mathbf{1}$	$\varepsilon\mathbf{k}$	$-\varepsilon\mathbf{j}$
$\mathbf{j}$	$\mathbf{j}$	$-\mathbf{k}$	$-\mathbf{1}$	$\mathbf{i}$	$\varepsilon\mathbf{j}$	$-\varepsilon\mathbf{k}$	$-\varepsilon\mathbf{1}$	$\varepsilon\mathbf{i}$
$\mathbf{k}$	$\mathbf{k}$	$\mathbf{j}$	$-\mathbf{i}$	$-\mathbf{1}$	$\varepsilon\mathbf{k}$	$\varepsilon\mathbf{j}$	$-\varepsilon\mathbf{i}$	$-\varepsilon\mathbf{1}$
$\varepsilon\mathbf{1}$	ε	$\varepsilon\mathbf{i}$	$\varepsilon\mathbf{j}$	$\varepsilon\mathbf{k}$	0	0	0	0
$\varepsilon\mathbf{i}$	$\varepsilon\mathbf{i}$	$-\varepsilon\mathbf{1}$	$\varepsilon\mathbf{k}$	$-\varepsilon\mathbf{j}$	0	0	0	0
$\varepsilon\mathbf{j}$	$\varepsilon\mathbf{j}$	$-\varepsilon\mathbf{k}$	$-\varepsilon\mathbf{1}$	$\varepsilon\mathbf{i}$	0	0	0	0
$\varepsilon\mathbf{k}$	$\varepsilon\mathbf{k}$	$\varepsilon\mathbf{j}$	$-\varepsilon\mathbf{i}$	$-\varepsilon\mathbf{1}$	0	0	0	0

Dual quaternions know two types of conjugation. The *conjugate quaternion* and the *conjugate dual quaternion* of a dual quaternion $Q = x_0 + \varepsilon y_0 + \mathbf{x} + \varepsilon\mathbf{y}$ are defined as

$$\overline{Q} = x_0 + \varepsilon y_0 - \mathbf{x} - \varepsilon\mathbf{y} \quad \text{and} \quad Q_e = x_0 - \varepsilon y_0 + \mathbf{x} - \varepsilon\mathbf{y}, \tag{2.24}$$

respectively. The norm of a dual quaternion is

$$\|Q\| = \sqrt{Q\overline{Q}}. \tag{2.25}$$

With these definitions, the equation $\mathbf{b} = \mathbf{X} \cdot \mathbf{a}$ where $\mathbf{X}$ is a matrix of the shape (2.3) can be written as

$$\mathbf{b} = (\varepsilon \mathbf{a})_e \star X \star \overline{\mathbf{a}} \tag{2.26}$$

where $X = \mathbf{x} + \varepsilon \mathbf{y}$, $\|X\| = 1$, $\mathbf{x} = (x_0, \ldots, x_3)^T$, $\mathbf{y} = (y_0, \ldots, y_3)^T$, and $\mathbf{x} \cdot \mathbf{y} = 0$. The last condition is precisely the Study condition (2.5).

In other words, Euclidean displacements can also be described by unit dual quaternions that satisfy the Study condition and kinematic mapping maps a Euclidean displacement to the corresponding unit dual quaternion. The algebra of dual quaternions provides a convenient way of computing in Study coordinates (see for example [23, Chapter 9]).

2.5. Geometry of the Study quadric. In Subsection 2.2 our topic was the geometry of the Study quadric S induced by coordinate changes in the fixed and in the moving space. Here we study the projective properties of S as hyper-quadric of seven dimensional projective space P^7. Our description follows [23, Section 11.2].

Lines in the Study quadric S correspond either to a one parameter set of rotations or to a one parameter set of translations. Lines through the identity ($[1 : 0 : \cdots : 0]$) correspond to one-parameter subgroups of SE(3) and are either rotation or translation subgroups.

The maximal subspaces of S are of dimension three ("3-planes"). More precisely, S is swept by two six dimensional families of 3-planes, called the *A-planes* and the *B-planes*. The A-planes and the B-planes are translates of the A-planes and the B-planes passing through the identity. Those 3-planes passing through the identity are the three dimensional subgroups of SE(3). They can be identified with SO(3) (the group of pure rotations) and with SE(2) (the group of planar Euclidean transformations). It is important to note that the exceptional three-space E, defined by $x_0 = x_1 = x_2 = x_3 = 0$, is an A-plane. The intersection of two A-planes or two B-planes is either empty or a one dimensional subspace. The intersection of an A-plane and a B-plane is either a point or a two dimensional plane. Whether an A-plane corresponds to SO(3) or SE(2) just depends on the intersection of the plane with E. In case of a point intersection the A-planes correspond to SO(3) and its translates; A-planes having line intersection with E correspond to SE(2) and its translates. General B-planes correspond to rotations about the axes in a plane, composed with a fixed displacement. The only B-plane that intersects the exceptional generator in a plane corresponds to the subgroup of all translations. All these cases belong of course to interesting kinematic configurations, but it would be beyond the scope of this paper to discuss all the possibilities. From algebraic point of view most attention has to be paid to the exceptional generator E because points in this space do not correspond to valid transformations in the pre-image space.

3. Mechanism theory. We start this section with a brief definition of basic concepts in mechanism science. Our terminology follows that of

TABLE 1
Important joint types.

Name	Abbr.	Dof	Relative motion
revolute	R	1	rotation about fixed axis
prismatic	P	1	translation in fixed direction
cylindrical	C	2	rotation about and translation along fixed axis
helical	H	1	rotation about and translation along fixed axis, linear relation between translation distance and rotation angle
spherical	S	3	rotation about axes through fixed point

[28, Section 1.2]. The fundamental object in computational kinematics is a *mechanism*. This is an object consisting of several links that are connected by joints.

A *link* is a collection of mechanical parts such that no relative motion between the individual members can occur. A *joint* is a connection between two links. It restricts the relative motion that is possible between the two links. The joints can be classified according to the nature of this restriction. The number of free parameters to describe this relative motion is called the *degree of freedom* of the joint. A listing of the most important joint types, their usual abbreviation, their degree of freedom, and a short description is given in Table 1.

It is important to note that the kinematic image of all joints in Table 1 – with exception of the helical joint – is an algebraic variety. We restrict ourselves to algebraic joints only. Luckily helical joints are of little relevance in practice.

A collection of links that are connected by joints is called a *kinematic chain*. A kinematic chain can be represented by a graph [28, Section 7.3.2] where the links are the vertices and the joints are the edges. In a *closed-loop kinematic chain* every link is connected to every other link by at least two paths, in an *open-loop kinematic chain* every link is connected to every other link by exactly one path. Of course there are also hybrid versions.

Finally, a *mechanism* is a kinematic chain where one of the links (the *base*) is fixed to the ground or, in mathematical terminology, to a base frame coordinate system. The remaining links are grouped into input links and output links. Input links are actuated and move with respect to the fixed link and the output links perform an according motion.

An example of a mechanism, a so-called planar 3R-linkage, is depicted in Figure 2. It consists of three revolute joints **a**, **b**, and **c**, connected by links of constant lengths. The joint **a** is fixed to the ground, **b** and **c** can move along the paths imposed by the links **ab** and **bc**, respectively. Attached to the last joint is the end effector tool. Typically, one is interested

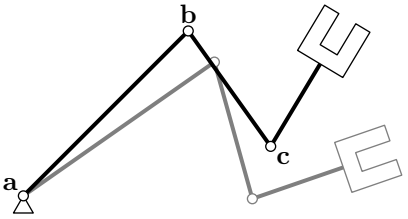


FIG. 2. *3R-linkage.*

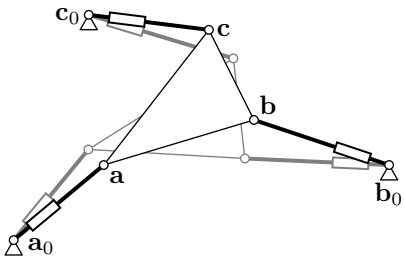


FIG. 3. *3RPR-platform.*

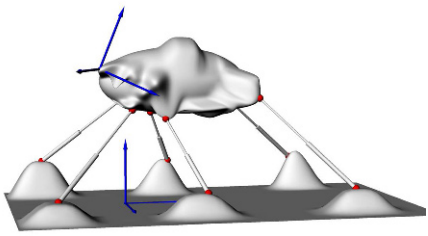


FIG. 4. *A general Stewart-Gough platform.*

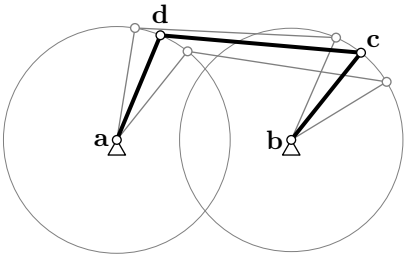


FIG. 5. *A planar four-bar mechanism.*

in the motion of the end effector tool with respect to the base. The set of all poses the end effector can attain is called the mechanism's *workspace*. The workspace is a subset of the Study quadric (or of planar or spherical kinematic image space).

In Figure 3 we see a planar 3RPR-platform. It consists of three legs, each composed of a revolute joint (R), a prismatic joint (P) and a further revolute joint (R). Three revolute joints ($\mathbf{a}_0$, $\mathbf{b}_0$, $\mathbf{c}_0$) are fixed to the ground, three of them ($\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$) are attached to the end effector frame. This mechanism is actuated by changing the lengths of the prismatic joints. It has a three dimensional workspace, that is represented by a three dimensional variety on the Study quadric.

The spatial counterpart to a 3RPR-platform is known as Stewart-Gough platform. It consists of six legs and each leg is composed of a spherical, a prismatic and a spherical joint (Figure 4). Between any two corresponding spherical joints a prismatic joint is inserted. The spherical joints can rotate freely about their center, the prismatic joints can extend or shrink in one direction (the direction of the link in our case).

Figure 5 depicts a planar four-bar linkage, a further common linkage type. It consists of four revolute joints connected by bars of fixed lengths. Two revolute joints $\mathbf{a}$, $\mathbf{b}$ are fixed to the base frame, the two remaining joints $\mathbf{c}$, $\mathbf{d}$ are attached to the end effector frame. The "missing" fourth

bar is the ideal connection between **a** and **b**. The four-bar motion depends only on one free parameter, the rotation of the driving crank. We say that the mechanism has *one degree of freedom*. This is in contrast to 3R- and 3RPR-manipulators which have three degrees of freedom and can, at least theoretically, generate the complete group of planar Euclidean displacements.

3.1. Serial and parallel manipulators. There is a fundamental difference between the 3R-linkage of Figure 2 and the 3RPR-platform of Figure 3. In case of the 3R-linkage, the joints are connected in a series while in case of the 3RPR-platform each two RPR-legs form a loop with end effector and base. The 3R-linkage is called a *serial manipulator* while the 3RPR-platform is called a *parallel manipulator*.

Parallel manipulators offer several advantages over serial ones: higher stiffness, higher payload capacity, lower inertia (see [28, p. 21]). On the other hand, serial manipulators are simpler and their workspace is usually larger. As far as computational kinematics is concerned, they exhibit a significantly different behavior in direct and inverse kinematics (see Subsection 4.2).

4. Constraint varieties. In this section we demonstrate how kinematic mapping can be used to translate mechanisms to algebraic varieties in P^7. These varieties describe the possible configurations of the mechanism and are called *constraint varieties*. We start by computing the kinematic images of fundamental building blocks of mechanisms (see Table 1). Then we demonstrate how to combine these elements in order to describe more complex mechanisms. The corresponding algebraic operations involve intersection of varieties and implicitization.

4.1. Kinematic image of elementary joints.

4.1.1. Revolute joints. A parametrized representation of the kinematic image of a revolute joint has already been computed in Examples 1 and 2. It is a straight line and computing its algebraic equations is elementary. Still, we will show how to carry out these computations explicitly because this demonstrates a general procedure for obtaining constraint varieties. Consider the kinematic image (2.10). It is given in a normal form and we see that it is described by six linear equations

$$\begin{aligned} &H_1(\mathbf{x})\colon x_1 = 0,\ H_2(\mathbf{x})\colon x_2 = 0,\ K_0(\mathbf{x})\colon y_0 = 0,\\ &K_1(\mathbf{x})\colon y_1 = 0,\ K_2(\mathbf{x})\colon y_2 = 0,\ K_3(\mathbf{x})\colon y_3 = 0. \end{aligned} \tag{4.1}$$

In order to obtain the constraint variety of a revolute joint in general position we have to transform the hyperplanes $H_i(\mathbf{x})$, $K_j(\mathbf{x})$ via the projective transformation $\mathbf{T}_f$. This is done by substituting $\mathbf{T}_f^{-1}\mathbf{x}$ for $\mathbf{x}$. The new equations are

$$H_i'(\mathbf{x}) = H_i(\mathbf{T}_f^{-1}\mathbf{x}), \quad K_j'(\mathbf{x}) = K_j(\mathbf{T}_f^{-1}\mathbf{x}). \tag{4.2}$$

This procedure not only works for linear equations but for algebraic equations of arbitrary degree. Changes of coordinates in the moving frame are performed by using $\mathbf{T}_m$ instead of $\mathbf{T}_f$.

4.1.2. Prismatic joints. The kinematic image of a prismatic joint can be computed in the same way as a revolute joint. We start with a matrix describing a translation in a simple home position:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ t & 0 & 0 & 1 \end{bmatrix} \tag{4.3}$$

The matrix (4.3) describes a translation in direction of the z-axis by the vector $[0, 0, t]^T$. Its kinematic image

$$\mathbf{t} = [2 : 0 : 0 : 0 : 0 : 0 : 0 : -t]^T \tag{4.4}$$

is again a straight line on the Study quadric, parameterized by the translation distance t (compare Exercise 2.1). Note, however, that for $t \to \infty$ we obtain a point of the absolute generator space E. This is a geometric property in the sense that it is not affected by changes of the moving or the base frame, and it distinguishes the images of rotations and translations. The hyperplane equations describing (4.4) are:

$$\begin{aligned} H_1(\mathbf{x})\colon x_1 = 0,\ H_2(\mathbf{x})\colon x_2 = 0,\ H_3(\mathbf{x})\colon x_3 = 0, \\ K_0(\mathbf{x})\colon y_0 = 0,\ K_1(\mathbf{x})\colon y_1 = 0,\ K_2(\mathbf{x})\colon y_2 = 0, \end{aligned} \tag{4.5}$$

and the transformed equations read again

$$H_i'(\mathbf{x}) = H_i(\mathbf{T}_f^{-1}\mathbf{x}), \quad K_j'(\mathbf{x}) = K_j(\mathbf{T}_f^{-1}\mathbf{x}). \tag{4.6}$$

4.1.3. Concatenation of joints. The movement of many industrial robots can be described as a composition of six consecutive rotations about axes in space. These robots are called 6R-robots and their degree of freedom is six. Here, we compute the constraint variety to the composition of two consecutive rotations (a "2R-chain").

- We describe the rotation of the second axis A_1 in its home position, where it coincides with the z-axis. This is the set (4.1) of algebraic equations $H_i(\mathbf{x})$, $K_j(\mathbf{x})$.
- Now we consider the relative displacement from the first axis A_0 to A_1 (Equation (2.16) or (2.17)). Transforming the second axes according to (4.2) we obtain equations $H_i'(\mathbf{x})$, $K_j'(\mathbf{x})$. They describe a configuration in space where A_0 is in its home position.
- The rotation about A_0 is described by (2.10), or after algebraization, by the rational Study vector

$$[1 : 0 : 0 : t : 0 : 0 : 0 : 0]^T. \tag{4.7}$$

Using the entries of this vector, we construct the matrix $\mathbf{T}_f = \mathbf{T}_f(t)$ and transform the equations $H_i'(\mathbf{x})$, $K_j'(\mathbf{x})$. The resulting hyperplane equations depend on t and constitute a hybrid representation of our 2R-chain. Eliminating the parameter t (see [5, Chapter 3, §3]) we finally obtain a number of algebraic equations $L_i(\mathbf{x})$, describing the constraint variety of a 2R-chain.

The extension of this procedure to an arbitrary number of revolute or prismatic joints is straightforward.

EXERCISE 4.1. *Compute the constraint variety of the 2R-chain with axes $A_0\colon x = y = 0$ and $A_1\colon x = z$, $y = 1$. (Hint: Compute at first the feet of the common normal of A_0 and A_1, then the Denavit-Hartenberg parameters d, a and α; see Figure 1). Show that this variety is the intersection of the Study quadric and a three dimensional space (this and the converse are generally true, see [23, p. 256]).*

EXERCISE 4.2. *Compute the constraint variety of the PR-chain where both, the revolute and the translation axes coincide with the z-axes. This is the constraint variety of a cylindrical joint in home position.*

4.1.4. Path constraints. So far, we have demonstrated how to operate on a "joint level" in order to compute constraint varieties. This is suitable for serial manipulators. When describing parallel manipulators it is often favorable to start with a "path constraint". Consider, for example, the four-bar mechanism of Figure 5. The coupler motion is completely defined by the condition that the two points have circular trajectories. Every "circle constraint" translates into a constraint surface in the quasielliptic space of planar Euclidean displacements. The kinematic pre-image of their intersection curve is the four-bar motion.

As an example we compute the algebraic equation of the surface of all planar displacements, such that the point $(a, b)^T$ moves on a circle with center $(\xi, \eta)^T$ and of radius ϱ. Using the matrix $\mathbf{X}$ of Equation (2.18) the circle constraint reads

$$\|\mathbf{X} \cdot (1, a, b)^T - (1, \xi, \eta)^T\|^2 - \varrho^2 = 0. \tag{4.8}$$

This is equivalent to a homogeneous polynomial of degree two in x_0, x_3, y_1, and y_2. Hence the circle constraint surface is a quadric surface in P^3.

EXERCISE 4.3. *Show that the circle constraint surface contains the absolute points $[0 : 0 : 1 : \pm\mathrm{i}]^T$ and is tangent to the planes $x_0 \pm \mathrm{i}x_3 = 0$.*

EXERCISE 4.4. *Compute the circle constraint surface of spherical kinematics and the sphere constraint surface of spatial kinematics (they are quadratic as well). Show that both constraint surfaces contain the exceptional quadric F.*

4.2. Mechanism analysis. The topic of mechanism analysis is the investigation of properties a certain mechanism exhibits. Thereby, the mechanism type and its dimensions are known. Questions of interest concern

the relation of joint parameters to the position and orientation of the end effector, the topology and size of the workspace, and its singular positions (singular in kinematic sense, not in the sense of algebraic geometry).

4.2.1. Direct and inverse kinematics. Direct and inverse kinematics are two basis tasks of mechanism analysis. In the direct kinematics problem the state of the input joints is known and the displacement of the end effector frame is sought. The inverse kinematics problem asks for the state of the joints when the position and orientation of the end effector frame are known.

Usually the direct kinematics problem is relatively easy for serial manipulators but often difficult for parallel manipulators. Conversely, the inverse kinematics problem is usually simple for parallel manipulators and often complicated for serial manipulators.

Consider for example an 6R serial chain. If the rotation angles of the individual joints are known, computing the end effector frame is just a matter of multiplying consecutive transformation matrices [19, Section 4.4] (direct kinematics). Conversely, it is not obvious at all how to choose the joint angles such that the end effector attains a certain specified pose (inverse kinematics). In the 1970s the inverse kinematics problem of a 6R chain was called as the "Mount Everest Problem" of kinematics [7] but since then efficient methods for computing its 16 solution sets were developed. See [19, Section 2] for an overview.

On the other hand, computing the leg lengths of a Stewart-Gough platform (SGP) is trivial when the position of the moving platform – and hence also the position of the anchor points – in space is given (inverse kinematics). The direct kinematics problem of finding the possible positions to a given sequence of leg lengths is difficult. It amounts to computing the intersection points of six sphere constraint surfaces and the Study quadric and has 40 solutions over $\mathbb{C}$. We give a short sketch of the solution algorithm. If a point of the moving system is constrained to remain on a sphere we obtain the following constraint equation:

$$\begin{aligned} h\colon\ & R(x_0^2+x_1^2+x_2^2+x_3^2)+4(y_0^2+y_1^2+y_2^2+y_3^2)-2x_0^2(Aa+Bb+Cc)\\ & +2x_1^2(-Aa+Bb+Cc)+2x_2^2(Aa-Bb-Cc)+2x_3^2(Aa+Bb+Cc)\\ & +2x_3^2(Aa+Bb-Cc)+4[x_0x_1(Bc-Cb)+x_0x_2(Ca-Ac)\\ & +x_0x_3(Ab-Ba)-x_1x_2(Ab+Ba)-x_1x_3(Ac+Ca)\\ & -x_2x_3(Bc+Cb)+(x_0y_1-y_0x_1)(A-a)+(x_0y_2-y_0x_2)(B-b)\\ & +(x_0y_3-y_0x_3)(C-c)+(x_1y_2-y_1x_2)(C+c)\\ & -(x_1y_3-y_1x_3)(B+b)+(x_2y_3-y_2x_3)(A+a)]=0, \end{aligned} \tag{4.9}$$

the expanded three dimensional version of (4.8) and part of the solution to Exercise 4.4. In this equation we have seven design constants: the length of the leg is encoded in R, the coordinates of the sphere center in the base

are A, B, C and the the center of the spherical joint on the platform are a, b, c. The direct kinematics problem is now transformed into an algebraic intersection problem of seven quadratic varieties (six constraint equations and the Study quadric S). The count of the number of solutions is not as easy because a simple check of the equation (4.9) shows that each of the equations contains the absolute quadric F in the exceptional generator $E\colon x_0 = x_1 = x_2 = x_3 = 0$. A proof for the existence of 40 solutions in the allowed part of S can be found in [29] or [23]. The solution algorithm is straightforward: take differences of the constraint equations, which are linear in y_i, solve for the y_i, substitute in the remaining equations. Three essentially different equations remain. They are in the general case of degree $(8, 4, 4)$. Using resultants or Gröbner bases the univariate polynomial of degree 40 can be computed and solved numerically.

EXERCISE 4.5. *Show that the direct kinematics problem of a planar 3RPR-platform of Figure 3 has, in general, six solutions over* $\mathbb{C}$. *(Hint: Use the fact that the circle constraint surface* (4.8) *is a quadric that, according to Exercise 4.4, contains the absolute points.)*

Also the other discussed kinematic problems are related to algebraic geometry by the fact that they give rise to systems of algebraic equations. One is interested in efficient numeric or symbolic algorithms for their solution, and in the maximum number of real solutions. Typically, like in the direct kinematics of SGP these systems are sparse and offer a lot of geometric structure that can be used to simplify computations.

4.2.2. Algebraic definition of degrees of freedom. There is a long history in defining and computing the degree of freedom of a mechanical system. Historically most of the developed formulas determine the topological structure and fail whenever special design parameters cause anomalies. Our informal definition at the beginning of Section 3 is an example of that. Exceptional, pathological or overconstrained mechanisms need special treatment. An overview of most of the classical concepts starting with Euler's formula up to the most recent developments can be found in [1].

Within the setting of algebraic geometry and the theory developed in this chapter it is natural to define the *degree of freedom* of a mechanism as the Hilbert dimension of the algebraic variety associated with the mechanical device. Caution has to be taken with respect to reality of the variety, components of different dimensions and parts of the variety that are completely contained in the exceptional generator. This will be seen explicitely in the following example.

Applying, for example, the above definition to the direct kinematics of the Stewart-Gough platform we obtain: The Hilbert dimension of the ideal spanned by the six sphere constraint equations (4.9) and the Study condition (2.5) is two because every sphere constraint variety contains the exceptional quadric F. But it is very well known that the direct kinematics

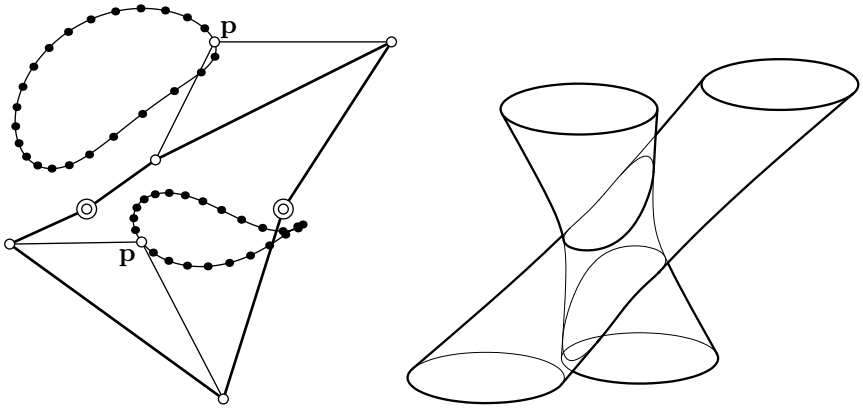


FIG. 6. *Two assembly modes of a four-bar and kinematic image.*

of a general Stewart-Gough platform has 40 discrete solutions. That is, for fixed leg lengths the degree of freedom is, in general, zero. Therefore, the algebraic variety should have at least one zero dimensional component. The dimensional problem can be overcome easily by adding a normalizing condition (either $x_0 = 1$ or $x_0^2 + x_1^2 + x_2^2 + x_3^2 = 1$) which removes the exceptional generator from the ideal.

For a Stewart-Gough platform having special design parameters the actual degree of freedom can be greater than zero. This interesting phenomenon will be the topic in Subsection 4.2.4.

4.2.3. Workspace topology. The *workspace* of a mechanism is defined as the union of all poses (position and orientation) the moving frame can attain. Its kinematic image is a real algebraic variety. The topology of this variety is an important property of a given mechanism.

Consider, as an example, the two manifestations of the four-bar mechanism in Figure 6. The four-bar has two *assembly modes*, that is, it can be assembled in two different states and one state cannot be reached from the other through a continuous series of four-bar displacements. This can be seen from the fact that the trajectory of a point $\mathbf{p}$ in the moving frame consists of two components. The kinematic image of the four bar motion is the intersection curve of two quadrics. So from an algebraic point of view, it is a complex curve, possibly having two disconnected real components. From the practical view of a mechanical engineer, it could be a mechansim having two different assembly modes.

Computing the workspace topology is an important task in theoretical kinematics. Via kinematic mapping this relates to computing the topology of algebraic varieties (see for example [10]). Even more important are methods for deciding whether two given poses lie in different assembly modes. In the language of real algebraic geometry this question can be

formulated as follows. Given are two points $\mathbf{p}$, $\mathbf{q}$ of an algebraic variety V. Are $\mathbf{p}$ and $\mathbf{q}$ contained in two disconnected components of V or not? An algorithmic solution for the case $\dim V = 1$ is presented in [16]. A simple test for four-bars is given in [22].

4.2.4. Mechanism singularities. The singular configurations of a mechanism are an important topic in mechanism analysis. The precise definition and classification of mechanism singularities is far beyond the scope of this article (compare [31] and the references therein). In particular, singular configurations of a mechanism do not necessarily correspond to singularities of the mechanism's constraint varieties. To obtain for example a formal definition of singularity of parallel mechanisms we can follow the exposition in [5] and apply the results therein to the constraint varieties of the mechanism. Let $V \in k^n$ be a constraint variety and let $p = [p_0, \ldots, p_7]^T$ be a point on V. The *tangent space* of V at p, denoted $T_p(V)$, is the variety

$$T_P(V) = \mathbf{V}(d_p(f) \colon f \subset \mathbf{I}(\mathbf{V})) \tag{4.10}$$

of linear forms $d_p(f)$ of all polynomials contained in the ideal $\mathbf{I}(\mathbf{V})$ in point p (see [5], page 486). With this definition we can immediately link the tangent space to the local degree of freedom of the mechanism: The local degree of freedom is defined as $\dim T_p(V)$. Note that it can be different from the global degree of freedom. Computationally the differentials are to be taken with respect to the Study parameters x_i, y_i. In kinematics these differentials are collected in the *Jacobian matrix* of the manipulator

$$\mathbf{J}(f_j) = \left(\frac{\partial f_j}{x_i}, \frac{\partial f_j}{y_i} \right), \tag{4.11}$$

where f_j are polynomials describing the constraints, the Study condition, and a normalizing condition. The normalizing condition has to be added to avoid dimensional problems coming from the exceptional generator E. In a nonsingular position of the mechanism the Jacobian $\mathbf{J}$ will have maximal rank. A singular position is characterized by rank deficiency of $\mathbf{J}$ and, if V is reduced, the defect is directly related to the local degree of freedom.

It should be noted that singularity of mechanisms has different meaning when applied to serial or parallel robots. In case of a serial manipulator singularity means loss of mobility, whereas in case of parallel manipulators singularity means gain of mobility. Singular or near-singular configurations have to be avoided because of unpredictable behavior of the platform, because its resistance towards forces in certain directions becomes very weak, and because the effect to manufacturing tolerances increases. The kinematic image of all singular configurations constitutes the manipulator's singularity surface.

There is a vast literature on singularities of mechanisms. In this article we confine ourselves to discussing singularities by means of a few examples. Some of them will show that the tangent space of the constraint varieties

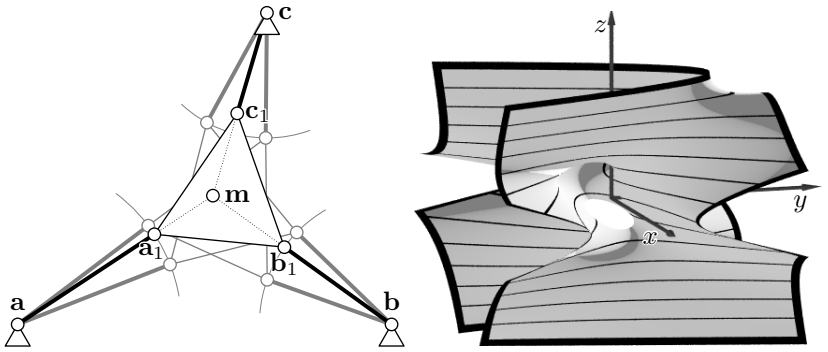


FIG. 7. *Singular configuration of a planar 3RPR-mechanism and singularity surface.*

does not always has to be computed. In these cases geometric consideration can replace the computation.

Planar 3RPR-platforms. A planar 3RPR-platform (Figure 3) is in a singular configuration, if the straight lines determined by the axes of its three legs intersect in a common point. It can be shown that this corresponds to a configuration where two solutions of the direct kinematics problem coincide. Therefore, the mechanisms behavior in a singular configuration is unpredictable. This is illustrated in Figure 7. The manipulator is in a singular configuration because the three legs meet in a common point **m**. Suppose now that we want to actuate the manipulator by changing the length of the leg through $\mathbf{c}_0$ while keeping fixed the remaining two leg lengths. This inverse kinematics problem has two solutions in the vicinity of the singular configurations.

The singularity surface Φ of a planar 3RPR-manipulator is depicted in Figure 7. Its equation is found by writing the positions of $\mathbf{a}_1$, $\mathbf{b}_1$, and $\mathbf{c}_1$ in general form (using (2.18)) and expanding the concurrency condition of three lines. Knowledge of geometric properties of Φ is helpful for singularity avoiding motion planing.

Stewart-Gough platforms. A SGP parallel manipulator is in a singular configuration when the Jacobian matrix (4.11) is rank deficient. There is again a simple geometric explanation for the singularity. Consider the axes of the legs of the manipulator. They are linearly dependent if they lie in a linear complex, a linear congruence or are lines on a quadric surface (see [20] for a definition of these concepts). There are more degenerate cases, which will not be mentioned here, but all cases are treated exhaustively in [18]. Because of the condition $\det \mathbf{J} = 0$ all singular positions of the manipulator belonging to rank deficiency 1 of $\mathbf{J}$ are on a degree 12 hyper-surface in P^7. Higher rank defect of $\mathbf{J}$ can be expressed by the vanishing of certain sub-determinants and corresponds to an algebraic variety as well. Little is known on that.

If $\det \mathbf{J} \equiv 0$, which means that the manipulator is singular independently of the position, then the manipulator is called *architectural singular*. One would expect that there are too many conditions that have to be fulfilled to allow this phenomenon. But surprisingly this is not the case. In [12] and [14] it is shown that for general SGP with arbitrarily distributed centers of the spherical joints architectural singularity is only possible for very degenerate designs. In the case of spherical joint centers being distributed in two planes four algebraic conditions are found which determine the locations of the anchor points in the two planes.

Self-motions of SGP occur when the mechanism moves without changing the leg length, that is with locked actuators. Algebraically this happens when the six constraint varieties and the Study quadric S determine at least a one dimensional ideal. The most famous example of this behavior is the Griffis-Duffy platform (Fig. 8).

Depending on the special design variables the constraint varieties determine various types of one dimensional ideals (see [13] and [24]). It also can happen that the ideal consist of different components having different dimensions. A simple example is the planar 3RPR-platform, when the base anchor points and the platform anchor points form two congruent triangles and the legs have the same length. The three constraint quadrics in the kinematic image space have a circle in common and four more points, of which two are the absolute points. The circle corresponds to the possible parallel-bar motion and the two points correspond to two rigid assembly modes. Self motions of platforms can be linked to an old and famous question in kinematics which was the topic of a competition of the French academy of science in 1904 (Prix Vaillant):

> Déterminer et étudier tous déplacements d'une figure invariable dans lesquels les differents points de la figure décrivent des courbes sphériques.
> (Determine and study all displacements of a rigid body in which distinct points of the body move on spherical paths.)

At the time posed this was a very difficult problem and not many mathematicians were able to give even partial answers. Of course, there are trivial cases: In a spherical motion all points move on a sphere. If two rigid bodies are connected by five rods of fixed length with spherical joints on both ends, one body can perform a one-parameter motion with five spherical trajectories. These and other examples were of course known when the prize question was posed. But the French academy wanted answers to further questions: Are there non trivial motions where all (or "many") points move on spheres? What are the displacements and of which type are the paths? Do such motions exist that have a higher degree of freedom? Among the eight submitted papers, two were awarded a prize, one by Émile Borel, the other by Raoul Bricard. Neither of them could give a complete classification of non trivial motions where all points run on spheres (it

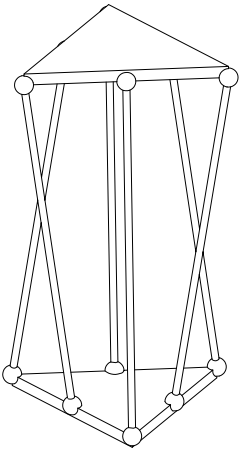


FIG. 8. *Griffis-Duffy platform.*

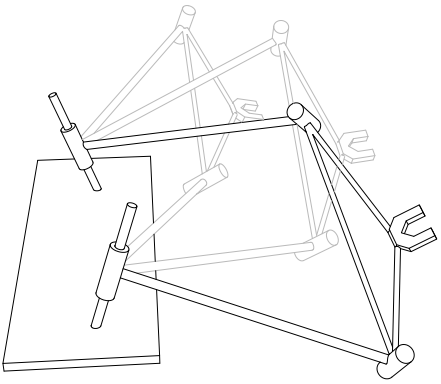


FIG. 9. *Bennett mechanism with three poses.*

is still missing) but they were able to describe numerous special cases. All of these examples are of relevance in robotics since they lead to self-motions of parallel manipulators of Stewart-Gough type (Section 4.2.2). More information and historical references on this academic competition can be found in [11].

4.3. Mechanism synthesis. Mechanism synthesis, as opposed to mechanism analysis, deals with mechanisms of yet unknown dimensions. The aim is to design a mechanism that is capable of performing a certain task. The complete theory of mechanism synthesis is a vast field [17]. Once the decision for a certain mechanism type is made, one has to determine the mechanism dimensions, its position in space and the position of the end effector tool with respect to the output link. The last step typically involves the solution of a system of algebraic equations. Often it is necessary to test the synthesized mechanism for violations of constraints that cannot be easily incorporated into the equation system (for example assembly mode defects, see Subsection 4.2.3).

Kinematic mapping based synthesis methods are most suitable for *motion generation*. This means that the synthesized mechanism's end effector has to attain certain poses, either exact or approximate. Accordingly, one distinguishes between exact and approximate synthesis. Approximate synthesis in kinematic image spaces in the sense of this text is problematic because of difficulties to define a meaningful distance between two displacements. Exact synthesis translates to the problem of interpolation by certain families of algebraic varieties. In principle, the design equations can be setup exactly as presented in Subsection 4.1. The only difference is that the design parameters of the mechanism, for example the Denavit-Hartenberg parameters of the relative position of two revolute axes, are unknown, while

the coordinates of the kinematic image are eliminated by substituting the coordinates of the prescribed poses.

As an example we show the design of a Bennett-mechanism from three given poses of the coupler system (Figure 9). A Bennett mechanism is a spatial closed 4R-loop mechanism. Generally a closed 4R would be rigid, but special design conditions found by Bennett in 1903 allow a one parameter motion [2]. The synthesis problem of a Bennett mechanism is completely determined when three poses of the coupler system are known. Until recently it was believed that the synthesis problem is of degree three. But a simple consideration based on 4.1 shows that it is linear: opening the closed 4R chain at the coupler link yields two 2R open serial chains. In the kinematic image space both chains are three-planes. Two three planes in P^7 generally have no intersection, which complies with the remark that a general 4R is rigid. But in the Bennett case the intersection must be a plane which on the other hand is completely defined by the three points. The kinematic image of a Bennett motion is a conic section.

REFERENCES

[1] R. Alizade, C. Bayram, and E. Gezgin, *Structural synthesis of serial platform manipulators*, Mechanism and Machine Theory, **42** (2007), pp. 580–599.

[2] G.T. Bennett, *A new mechanism*, Engineering, **76** (1903), pp. 777–778.

[3] W. Blaschke, *Euklidische Kinematik und nichteuklidische Geometrie*, Zeitschr. Math. Phys., **60** (1911), pp. 61–91; 203–204.

[4] O. Bottema and B. Roth, *Theoretical Kinematics*, Dover Publications, 1990.

[5] D.A. Cox, J.B. Little, and D. O'Shea, *Ideals, Varieties and Algorithms*, Springer, third ed., 2007.

[6] H.S.M. Coxeter, *Non-Euclidean Geometry*, Math. Assoc. Amer., 6th ed., 1988.

[7] F. Freudenstein, *Kinematics: Past, present and future*, Mechanism and Machine Theory, **8** (1973), pp. 151–160.

[8] G.H. Golub and C.F. Van Loan, *Matrix Computations*, Baltimore: Johns Hopkins University Press, third ed., 1996.

[9] J. Grünwald, *Ein Abbildungsprinzip, welches die ebene Geometrie und Kinematik mit der räumlichen Geometrie verknüpft*, Österreich. Akad. Wiss. Math.-Natur. Kl. S.-B. II, **80** (1911), pp. 677–741.

[10] A. Hatcher, *Algebraic Topology*, Cambridge University press, 2002.

[11] M.L. Husty, *E. Borel's and R. Bricard's papers on displacements with spherical paths and their relevance to self-motions of parallel manipulators*, in International Symposium on History of Machines and Mechanisms-Proceedings HMM 2000, M. Ceccarelli, ed., Kluwer Acad. Pub., 2000, pp. 163–172.

[12] M.L. Husty and A. Karger, *Architecture singular parallel manipulators and their self-motions*, in Advances in Robot Kinematics, J. Lenarcic and M.M. Stanisic, eds., Kluwer Acad. Pub., 2000, pp. 355–364.

[13] ———, *Self-motions of Griffis-Duffy type platforms*, in Proceedings of IEEE conference on Robotics and Automation (ICRA 2000), San Francisco, USA, 2000, pp. 7–12.

[14] ———, *Architecture singular planar Stewart-Gough platforms*, in Proceedings of the 10th workshop RAAD, Vienna, Austria, 2001, p. 6. CD-Rom Proceedings.

[15] M.L. Husty, A. Karger, H. Sachs, and W. Steinhilper, *Kinematik und Robotik*, Springer, Berlin, Heidelberg, New York, 1997.

[16] Y. Lu, D. Bates, A.J. Sommese, and C.W. Wampler, *Finding all real points of a complex curve*, in Proceedings of the Midwest Algebra, Geometry and Its Interactions Conference, Contemporary Mathematics, AMS, 2007, p. v. 448.
[17] J.M. McCarthy, *Geometric Design of Linkages*, Vol. 320 of Interdisciplinary Applied Mathematics, Springer, New York, 2000.
[18] J.-P. Merlet, *Singular configurations of parallel manipulators and Grassmann geometry*, Int. Journ. of Robotics Research, **8** (1992), pp. 150–162.
[19] M. Pfurner, *Analysis of spatial serial manipulators using kinematic mapping*, PhD thesis, University Innsbruck, 2006.
[20] H. Pottmann and J. Wallner, *Computational Line Geometry*, Springer, 2001.
[21] W. Rath, *Matrix groups and kinematics in projective spaces*, Abh. Math. Sem. Univ. Hamburg, **63** (1993), pp. 177–196.
[22] H.-P. Schröcker, M.L. Husty, and J.M. McCarthy, *Kinematic mapping based assembly mode evaluation of planar four-bar mechanisms*, ASME J. Mechanical Design, **129** (2007), pp. 924–929.
[23] J.M. Selig, *Geometric Fundamentals of Robotics*, Monographs in Computer Science, Springer, New York, 2005.
[24] A. Sommese, J. Verschelde, and C. Wampler, *Advances in polynomial continuation for solving problems in kinematics*, ASME J. Mechanical Design, **126** (2004), pp. 262–268.
[25] A. Sommese and C. Wampler, *The Numerical Solution of Systems of Polynomials Arising in Engineering and Science*, World Scientific, 2006.
[26] E. Study, *Von den Bewegungen und Umlegungen*, Math. Ann., **39** (1891), pp. 441–566.
[27] ———, *Geometrie der Dynamen*, B.G. Teubner, Leipzig, 1903.
[28] L.-W. Tsai, *Robot Analysis: The Mechanics of Serial and Parallel Manipulators*, John Wiley & Sons, Inc., 1999.
[29] C.W. Wampler, *Forward displacement analysis of general six-in-parallel SPS (Stewart) platform manipulators*, Mechanism and Machine Theory, **31** (1996), pp. 331–337.
[30] W. Wunderlich, *Ein vierdimensionales Abbildungsprinzip für ebene Bewegungen*, Z. Angew. Math. Mech., **66** (1986), pp. 421–428.
[31] D. Zlatanov, R.G. Fenton, and B. Benhabib, *Identification and classification of the singular configurations of mechanisms*, Mechanism and Machine Theory, **33** (1998), pp. 743–760.

RATIONAL OFFSET SURFACES AND THEIR MODELING APPLICATIONS

RIMVYDAS KRASAUSKAS* AND MARTIN PETERNELL†

Abstract. This survey discusses rational surfaces with rational offset surfaces in Euclidean 3-space. These surfaces can be characterized by possessing a field of rational unit normal vectors, and are called Pythagorean normal surfaces. The procedure of offsetting curves and surfaces is present in most modern 3d-modeling tools. Since piecewise polynomial and rational surfaces are the standard representation of parameterized surfaces in CAD systems, the rationality of offset surfaces plays an important role in geometric modeling. Simple examples show that considering surfaces as envelopes of their tangent planes is most fruitful in this context. The concept of Laguerre geometry combined with universal rational parametrizations helps to treat several different results in a uniform way. The rationality of the offsets of rational pipe surfaces, ruled surfaces and quadrics are a specialization of a result about the envelopes of one-parameter families of cones of revolution. Moreover a couple of new results are proved: the rationality of the envelope of a quadratic two-parameter family of spheres and the characterization of classes of Pythagorean normal surfaces of low parametrization degree.

Key words. rational surfaces, rational offsets, Pythagorean normal surfaces, LN surfaces, canal surfaces, Laguerre geometry, universal rational parametrization.

1. Introduction and the history of rational offset surfaces.

When modeling real world objects one not only uses surfaces but has to take into account the material thickness. Thus offsetting curves and surfaces is a frequently used tool and it is present in most of the 3d-geometry-modeling software nowadays. These systems typically represent parameterized curves and surfaces as B-splines or piecewise rational (NURBS) curves and surfaces. This motivated several researchers [10, 7–9, 22, 20, 21, 27, 28, 34, 35], just to name a few of them, to study rational curves and surfaces with rational offsets.

Given a parametric rational surface $\mathbf{f}(u, v)$ with unit normal vector field $\mathbf{n}(u, v)$, the offset surfaces at distance d can be represented parametrically by

$$\mathbf{f}_d(u, v) = \mathbf{f}(u, v) + d\mathbf{n}(u, v). \tag{1.1}$$

Because of the normalization of the normal vector $\mathbf{n}$, the offset surfaces of rational surfaces $\mathbf{f}(u, v)$ are typically non-rational. This also holds for curves. For instance the offsets of an ellipse are non-rational algebraic curves of degree eight. But even if the rational surface $\mathbf{f}(u, v)$ possesses rational offsets, the representation (1.1) is typically non-rational. This can already be realized for a parabola $\mathbf{c}(t) = (t, t^2)$, whose offsets are rational curves of degree six, but the parametrization $\mathbf{c}_d(t) = \mathbf{c}(t) + d\mathbf{n}(t)$ with

*Vilnius University, Lithuania (rimvydas.krasauskas@mif.vu.lt).
†Vienna University of Technology, Austria (peternell@geometrie.tuwien.ac.at),
Grant Austrian Science Fund FWF under project S92.

I.Z. Emiris et al. (eds.), *Nonlinear Computational Geometry*, The IMA Volumes in Mathematics and its Applications 151, DOI 10.1007/978-1-4419-0999-2_5,

$\mathbf{n}(t) = 1/\sqrt{1+4t^2}(-2t, 1)$ is non-rational. An appropriate reparametrization of the parabola is required to represent the offsets by rational parametrizations. Thus it is necessary to study this subject in more detail to be able to decide whether the offset surfaces of a given rational surface are rational and how to derive and construct rational parametrizations.

The rationality of a surface is determined by vanishing genus and second plurigenus. But the computation of these invariants is quite complex for surfaces given by parametric representations, such that determining the rationality of offset surfaces is difficult. Additionally we note that here we will denote a real surface as *rational* if and only if it admits a real rational parametrization. There exist real surfaces possessing rational (improper) parametrizations but their genus does not vanish, for instance offsets of ellipsoids.

The analogous questions for curves have been studied for a long time within the computer-aided-geometric-design community. Farouki [10, 7] introduced the notion of PH curves, see also the survey [8] and the recent book [9]. This term denotes *polynomial curves* $\mathbf{p}(t)$ with the property that the norm of the tangent vector $\dot{\mathbf{p}}(t)$ is polynomial. This implies that the *arc length* of $\mathbf{p}(t)$ is a *polynomial*. Setting $\dot{\mathbf{p}} = w(t)\left(u(t)^2 - v(t)^2, 2u(t)v(t)\right)$ with arbitrary polynomials $u(t), v(t)$ and $w(t)$, the norm $\|\dot{\mathbf{p}}\|$ as well as the norm of the normal vector equals the polynomial $w(t)(u(t)^2 + v(t)^2)$. Consequently the unit normal vector is rational. The concept of PH curves is also generalized to space curves, see e.g. [8].

Rational surfaces with rational offsets are more involved and the techniques used for curves do not apply to surfaces directly. An explicit representation of all rational surfaces with rational offsets has been given in [38]. Nevertheless it is not obvious how to decide the rationality of the offsets for particular classes of surfaces. It has been proved that rational pipe surfaces [28], rational ruled surfaces [42] and all regular quadrics [30] possess rational offsets. These statements can also be found in [35] as specializations of a more general result concerning envelopes of rational one-parameter families of cones of revolution.

Since any cone of revolution is the envelope of a one-parameter family of spheres as well as planes, the envelope is also generated by a two-parameter family of spheres. Using the affine space $\mathbb{R}^4$ as model of the four-dimensional manifold of spheres in Euclidean $\mathbb{R}^3$, the mentioned result reads: A rational ruled surface in the model space $\mathbb{R}^4$ represents a two-parameter family of spheres whose envelope surface as well as its offset surfaces possess rational parametrizations and these parametrizations can be constructed explicitly. This result is a general statement about a class of surfaces in $\mathbb{R}^4$ and their corresponding envelopes in $\mathbb{R}^3$.

The article describes the current status of research in the field of rational offset surfaces and it points to some new results and open questions. It will provide a short introduction to some theoretical tools which are necessary for their treatment. Section 2 provides a first and elementary

introduction to rational offset surfaces which are constructed using the Blaschke image of the space of planes. Section 3 deals with the special family of rational surfaces possessing a 'linear normal vector field'. In Section 4 we provide a theoretical investigation of the subject introducing to Euclidean Laguerre geometry, the geometry of oriented spheres and planes in $\mathbb{R}^3$ and its models. Section 5 gives rational parametrizations in full generality and Section 6 deals with several special families of rational offset surfaces. Section 7 is devoted to modeling applications and finally we conclude this article and discuss some open problems.

2. Different approaches to rational offset surfaces. We start with defining rational surfaces with rational offsets, discuss these surfaces as envelopes of spheres and planes and derive concepts to obtain an elegant approach to deal with these surfaces.

DEFINITION 2.1. *A surface F in $\mathbb{R}^3$ is a* Pythagorean normal surface *or* PN surface *if it possesses a rational parametrization $\mathbf{f}(u,v)$ and a rational unit normal vector field $\mathbf{n}(u,v)$ corresponding to $\mathbf{f}(u,v)$. The offset surface F_d of F at oriented distance d admits a rational parametrization $\mathbf{f}_d(u,v) = \mathbf{f}(u,v) + d\mathbf{n}(u,v)$.*

The rationality of a surface does not depend on a particular parametrization. Since we take a constructive viewpoint we like to construct parametrizations $\mathbf{f}(u,v)$ which directly lead to rational parametrizations $\mathbf{f}_d(u,v)$ of the offsets. The correspondence noted in Definition 2.1 means that $\mathbf{n}(u,v)$ is computed via normalizing the cross product $\mathbf{f}_u \times \mathbf{f}_v$,

$$\mathbf{n}(u,v) = \frac{1}{\|\mathbf{f}_u(u,v) \times \mathbf{f}_v(u,v)\|}\mathbf{f}_u(u,v) \times \mathbf{f}_v(u,v), \tag{2.1}$$

with $\mathbf{f}_u$ and $\mathbf{f}_v$ as partial derivatives of $\mathbf{f}$ with respect to u and v, respectively. If the norm $\|\mathbf{f}_u(u,v) \times \mathbf{f}_v(u,v)\|$ is a rational function, the parametrization (1.1) is a rational representation of the offsets F_d of F and is called *PN-parametrization*. But typically this norm involves square roots even for rational offset surfaces and appropriate reparametrizations have to be performed.

2.1. Offsets as envelopes of spheres and planes. Assuming $\mathbf{f}(u,v)$ is a PN-parametrization of a rational offset surface F, the tangent planes $E(u,v)$ of F admit the rational representation

$$E(u,v) : n_0(u,v) + \mathbf{n}(u,v) \cdot \mathbf{x} = 0, \text{ with } \|\mathbf{n}(u,v)\|^2 = 1. \tag{2.2}$$

The rational support function $n_0 = -\mathbf{n} \cdot \mathbf{f}$ expresses the oriented distance of the origin from E. Eqn. (2.2) interprets the surface F as the envelope of the two-parameter family of tangent planes $E(u,v)$. Likewise let the offset surface F_d of F be considered as the envelope of its tangent planes. Translating the planes E by the constant oriented distance d in direction

of the normal vector $\mathbf{n}$ results in the tangent planes E_d of the offset surface F_d, with

$$E_d(u,v) : n_0(u,v) - d + \mathbf{n}(u,v) \cdot \mathbf{x} = 0. \tag{2.3}$$

Given a parameterized surface F with representation $\mathbf{f}(u,v)$, the offset surface F_d at distance d can be considered as the envelope of a two-parameter family of spheres

$$S : (\mathbf{x} - \mathbf{f}(u,v)) \cdot (\mathbf{x} - \mathbf{f}(u,v)) - d^2 = 0, \tag{2.4}$$

of radius d which are centered at the surface F. According to the envelope condition an implicit equation of the offset surface F_d is obtained by eliminating the surface parameters u and v from the system of equations

$$S\colon (\mathbf{x}-\mathbf{f}) \cdot (\mathbf{x}-\mathbf{f}) - d^2 = 0, \quad \frac{\partial S}{\partial u}\colon (\mathbf{x}-\mathbf{f}) \cdot \mathbf{f}_u = 0, \quad \frac{\partial S}{\partial v}\colon (\mathbf{x}-\mathbf{f}) \cdot \mathbf{f}_v = 0. \tag{2.5}$$

Note that these two approaches are not equivalent. The first interpretation (2.3) yields one-sided offsets whereas the second one (2.5) results typically in two sheets of the offset surface at both sides of F. If we consider rational surfaces, both approaches might yield the same result if the original surface F is traced twice and thus both orientations of the normal vector field $\mathbf{n}(u,v)$ appear.

Now we focus on the interpretation of offset surfaces as the envelopes of their tangent planes (2.3). For this reason we introduce to the manifold of oriented planes of $\mathbb{R}^3$. Later we will see the close connection between oriented planes and spheres in Section 4.

2.2. The space of oriented planes and the Blaschke model. We consider the family of oriented planes $E : e_0 + \mathbf{e} \cdot \mathbf{x} = 0$, where $\mathbf{e}$ denotes the unit normal vector of E, similar to (2.2). The real numbers e_0, and $\mathbf{e} = (e_1, e_2, e_3)$ determine the oriented plane. We use $(e_1, e_2, e_3, e_0) \in \mathbb{R}^4$ with $\|\mathbf{e}\| = 1$ as coordinates of oriented planes. Denoting the family of planes of $\mathbb{R}^3$ by $\mathcal{E}$, this defines the *Blaschke mapping*

$$\beta : \mathcal{E} \to \mathbb{R}^4, \quad E : e_0 + \mathbf{e} \cdot \mathbf{x} = 0 \mapsto \beta(E) = (e_1, e_2, e_3, e_0), \tag{2.6}$$

which identifies oriented planes $E \in \mathcal{E}$ of $\mathbb{R}^3$ with the family of points $\beta(E) \in \mathbb{R}^4$. According to the normalization condition $\|\mathbf{e}\| = 1$ the image points $\beta(E)$ are contained in the quadratic cone

$$\mathcal{B} : y_1^2 + y_2^2 + y_3^2 = 1, \tag{2.7}$$

called the *Blaschke cylinder* $\mathcal{B}$. Here we use $y_1, \ldots, y_4$ as Cartesian coordinates in $\mathbb{R}^4$.

The intersections of $\mathcal{B}$ with 3-spaces $y_4 =$ constant are copies of the unit sphere $S^2 : x_1^2 + x_2^2 + x_3^2 = 1$ and $\mathcal{B}$ is a cylinder over S^2. Consider two parallel planes $E : e_0 + \mathbf{n} \cdot \mathbf{x} = 0$ and $F : f_0 + \mathbf{n} \cdot \mathbf{x} = 0$ with coinciding unit normal $\mathbf{n}$. Their image points $\beta(E)$ and $\beta(F)$ are contained in a generating line of $\mathcal{B}$.

2.3. The Blaschke image of a PN surface. The introduction of the Blaschke mapping provides a theoretical background for the study of PN surfaces and applies also to find rational parametrizations of these surfaces.

Let F be a PN surface with representation $\mathbf{f}(u,v)$ and tangent planes $E(u,v) : e_0(u,v) + \mathbf{e}(u,v) \cdot \mathbf{x} = 0$. The unit normal vector field $\mathbf{e}(u,v)$ is a rational parametrization of the unit sphere S^2. The Blaschke image $\beta(F) = \beta(E(u,v))$ is a rational surface in $\mathcal{B}$ with rational parametrization $\beta(E) = (e_1, e_2, e_3, e_0)(u,v)$.

THEOREM 2.1. *The Blaschke image $\beta(F)$ of a PN surface F is a rational surface in $\mathcal{B}$. Conversely, any rational surface in $\mathcal{B}$ is the Blaschke image of a PN surface.*

Given any rational two-dimensional surface S in the Blaschke cylinder $\mathcal{B}$, we have to consider it as image points of tangent planes. The envelope of this two-parameter family of planes corresponding to the surface $S \subset \mathcal{B}$ is a PN surface in $\mathbb{R}^3$. Let $\mathbf{S}(u,v) = (s_1, \ldots, s_4)(u,v)$ be a rational parametrization of S. The corresponding family of tangent planes is

$$T(u,v) : s_4(u,v) + \mathbf{s}(u,v) \cdot \mathbf{x} = 0,$$

where $\mathbf{s} = (s_1, s_2, s_3)$ denotes the unit normal vector of T. Computing the partial derivatives $T_u(u,v)$ and $T_v(u,v)$ of $T(u,v)$ gives a PN-parametrization $\mathbf{f}(u,v)$ of the surface F as solution of the system of linear equations

$$\begin{aligned} T(u,v) :\ & s_4(u,v) + \mathbf{s}(u,v) \cdot \mathbf{x} = 0, \\ T_u(u,v) :& s_{4u}(u,v) + \mathbf{s}_u(u,v) \cdot \mathbf{x} = 0, \\ T_v(u,v) :& s_{4v}(u,v) + \mathbf{s}_v(u,v) \cdot \mathbf{x} = 0. \end{aligned} \tag{2.8}$$

Until now we have restricted our interest to two-dimensional surfaces in $\mathcal{B}$. But what about curves? Consider a rational curve $C \subset \mathcal{B}$ with representation $\mathbf{C}(t) = (c_1, \ldots, c_4)(t)$. The corresponding surface $\beta^{-1}(C) = D$ is the envelope of a one-parameter family of planes and thus a developable PN surface. The generating lines $g(t)$ are obtained as intersections $T \cap T_t$ and are the solutions of the system of equations

$$\begin{aligned} T(t) :& c_4(t) + \mathbf{c}(t)^T \cdot \mathbf{x} = 0, \\ T_t(t) :& c_{4t}(t) + \mathbf{c}_t(t)^T \cdot \mathbf{x} = 0. \end{aligned} \tag{2.9}$$

By the way, the intersection $\mathbf{v}(t) = T \cap T_t \cap T_{tt}$ is in general the singular curve V of D. For special developable surfaces like cones and cylinders, V degenerates to a point or an ideal point, respectively. We summarize these results.

THEOREM 2.2. *The Blaschke image $\beta(F)$ of a developable PN surface F is a rational curve in $\mathcal{B}$. Conversely, any rational curve in $\mathcal{B}$ is the*

Blaschke image of a developable PN surface. The Theorems 2.1 and 2.2 can be found in a different form in [35]. There the stereographic projection of the Blaschke cylinder to a 3-space and the corresponding projections of the curves and surfaces $\beta(F)$ are investigated.

In the sequel we do not pay much attention to developable surfaces and thus we assume that the family of tangent planes of considered surfaces is two-dimensional, unless explicitly mentioned.

2.4. The Gaussian image of a PN surface. We consider a rational offset surface $F \subset \mathbb{R}^3$ whose tangent planes have the form $T(u,v) : e_0(u,v) + \mathbf{e}(u,v) \cdot \mathbf{x} = 0$, with $\|\mathbf{e}\| = 1$. The Blaschke image $\beta(F) = (e_1, e_2, e_3, e_0)$ of F consists of

- a rational parametrization $\mathbf{e}(u,v)$ of the unit sphere S^2, and
- the support function $e_0(u,v)$ of F.

The parametrization $\mathbf{e}(u,v)$ of the unit normal vector field of F is the *Gaussian image* of F. One can take apart the Blaschke image $\beta(F)$ and study the Gaussian image $\mathbf{e}(u,v)$ and the support function $e_0(u,v)$ separately. This constructive approach to PN surfaces is based on the study of rational parametrizations of the unit sphere through the Gauss map.

2.5. Rational parametrizations of the unit sphere via stereographic projection. The easiest way to construct rational parametrizations of the unit sphere S^2 is as follows. Let $a(u,v), b(u,v)$ and $c(u,v)$ be relatively prime bivariate polynomials. Then a rational parametrization $\mathbf{e} = (e_1, e_2, e_3)$ of S^2 is obtained by

$$e_1 = \frac{2ac}{n}, \quad e_2 = \frac{2bc}{n}, \quad e_3 = \frac{a^2 + b^2 - c^2}{n}, \text{ with } n = a^2 + b^2 + c^2. \quad (2.10)$$

This parametrization is a composition of a rational parametrization $\mathbf{x} = (\frac{a}{c}, \frac{b}{c})$ of $\mathbb{R}^2$ and the stereographic projection $\sigma : \mathbb{R}^2 \to S^2$ with center $(0,0,1)$

$$\sigma(\mathbf{x}) = \left(\frac{2x_1}{x_1^2 + x_2^2 + 1}, \frac{2x_2}{x_1^2 + x_2^2 + 1}, \frac{x_1^2 + x_2^2 - 1}{x_1^2 + x_2^2 + 1} \right).$$

The parametrization (2.10) is geometrically evident but has the drawback that it is dependent not only on the coordinate system but also on the choice of the center for the stereographic projection. To avoid this we will introduce universal rational parametrizations of S^2 in Section 5. This leads to universal parametrizations of the Blaschke cylinder $\mathcal{B} \subset \mathbb{R}^4$ which represent the most general approach to PN surfaces.

2.6. Rational parametrizations of PN surfaces in the simple form. Starting from the parametrization (2.10) of the unit sphere one can derive rational parametrizations of PN-surfaces as already developed

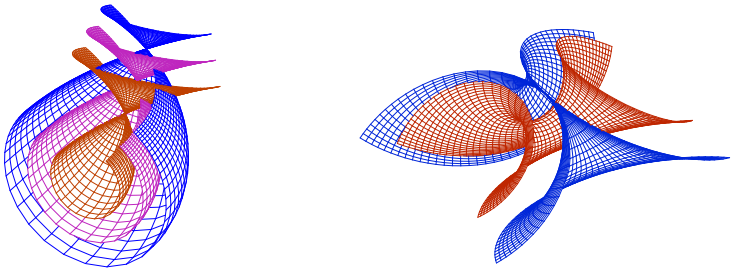


FIG. 1. *Parabolic Dupin cyclides and their offsets.*

in [38]. Prescribing a rational function $h(u,v) = f(u,v)/g(u,v)$, the tangent planes of a rational offset surface take the form

$$T(u,v) : h + \frac{2ac}{n}x_1 + \frac{2bc}{n}x_2 + \frac{a^2+b^2-c^2}{n}x_3 = 0.$$

Multiplying by the denominator n gives a polynomial representation of the tangent planes in the form

$$T(u,v) : fn + 2acgx_1 + 2bcgx_2 + g(a^2+b^2-c^2)x_3 = 0,$$

where f and g are polynomials without a common factor. This approach is best illustrated by an example.

EXAMPLE 1. *Assume that $a = u, b = v$ and $c = 1$. This leads to the standard form of a rational parametrization of S^2 by*

$$\left(\frac{2u}{u^2+v^2+1}, \frac{2v}{u^2+v^2+1}, \frac{u^2+v^2-1}{u^2+v^2+1}\right).$$

Further we assume $g = u^2+v^2+1$ and choose f as an arbitrary quadratic polynomial $q(u,v)$. The surfaces whose tangent planes can be parameterized by

$$T(u,v) : q(u,v) + 2ux_1 + 2vx_2 + (u^2+v^2-1)x_3 = 0 \tag{2.11}$$

are known as parabolic Dupin cyclides. These surfaces are of algebraic order 3 and form a family of surfaces which is closed under taking offsets (see Fig. 1). The real singularities of the surfaces and their offsets might be different.

3. Rational Surfaces with a linear normal vector field. A special class of rational offset surfaces is formed by the offsets of rational surfaces which possess a linear normal vector field, see [13]. Moreover it has been shown in [48] that the convolutions of these surfaces with all rational surfaces are again rational.

A rational surface F is called an *LN surface* if there exists a *rational* parametrization $\mathbf{s}(u,v)$ such that a normal vector field $\mathbf{n}(u,v)$ of S can be linearly parameterized as

$$\mathbf{n}(u,v) = \mathbf{p}u + \mathbf{q}v + \mathbf{r}, \text{ with } \mathbf{p},\mathbf{q},\mathbf{r} \in \mathbb{R}^3. \tag{3.1}$$

Note that the normal vector field $\mathbf{n}(u,v)$ is in general *not normalized* and *not oriented*. This is quite different from the previous approach to rational offset surfaces via PN surfaces in Section 2.3. Later on we show how this class of surfaces fits to the presented concept and how the Blaschke images $\beta(F)$ of LN surfaces look like.

In the following we assume that $\operatorname{rank}(\mathbf{p},\mathbf{q},\mathbf{r}) = 3$, the coordinate vectors $\mathbf{n}(u,v)$ parameterize points of an affine plane. This implies that the unit normal vectors of F parameterize a two-dimensional subset of S^2. Otherwise the corresponding surface F is either a cylinder or a plane. An appropriate choice of the coordinate system is $\mathbf{p} = (1,0,0)$, $\mathbf{q} = (0,1,0)$, and $\mathbf{r} = (0,0,1)$, and the normal vector becomes $\mathbf{n}(u,v) = (u,v,1)^T$, which we assume below. The tangent planes $T(u,v)$ of an LN surface F have the quite simple representation

$$T(u,v): \; h(u,v) + ux + vy + z = 0, \tag{3.2}$$

where $h(u,v)$ is a rational function. With respect to the chosen coordinate system, the tangent planes T are graphs of linear functions over the xy-plane. The representation (3.2) allows it to treat $(u,v,h(u,v))$ as affine coordinates of T. Using (U,V,W) as coordinate functions of planes, the *dual affine equation* of an LN surface F is $W = h(U,V)$. This representation says that the tangent planes of LN surfaces are graphs of rational functions.

This property has the following important consequence: For any vector $\mathbf{n} = (u,v,1)^T$ there exists a unique tangent plane $T(u,v)$ of F having $\mathbf{n}$ as normal vector and there exists exactly one point of contact of F and T. This *unique-tangent-plane-property* is the reason for the *rationality* of the convolution surfaces with any arbitrary rational surface, see [48]. Summarizing we obtain

COROLLARY 3.1. *The family of tangent planes $T(u,v)$ of an LN surface F can be represented in plane coordinates by the graph $(u,v,h(u,v))$ of a rational function h. Conversely, the graph of a rational function represents the tangent planes* (3.2) *of an LN surface. The convolution surface $F \star G$ of an LN surface F and any arbitrary rational surface G is a rational surface.*

3.1. The Blaschke image of LN surfaces. Since LN surfaces F are very special concerning their normal vectors and tangent planes, also their Blaschke image $\beta(F)$ is of a special kind. To obtain rational parametrizations of their Blaschke image $\beta(F)$ we have to reparameterize and normalize the unit normal vector field. Inserting the rational reparametrization

$$u = \frac{2s}{1-s^2-t^2}, \quad v = \frac{2t}{1-s^2-t^2} \tag{3.3}$$

into the representation (3.2) and normalizing the normal vector leads to

$$T(s,t): \frac{1-s^2-t^2}{1+s^2+t^2}h(s,t)+\frac{2s}{1+s^2+t^2}x+\frac{2t}{1+s^2+t^2}y+\frac{1-s^2-t^2}{1+s^2+t^2}z=0, \quad (3.4)$$

which exhibits that LN surfaces are rational offset surfaces (PN surfaces). The reparametrization (3.3) induces an orientation to the plane $T(s,t)$ as well as to the surface F. The substitution (3.3) is not one-to-one, but there exist parameters

$$s' = \frac{-s}{s^2+t^2}, \quad \text{and} \quad t' = \frac{-t}{s^2+t^2},$$

for which $u(s,t) = u(s',t')$ and $v(s,t) = v(s',t')$ holds. The normal vector $\mathbf{n}(u,v) = (u,v,1)$ corresponds to and is parallel to the two oriented unit normal vectors

$$\mathbf{n}(s,t) = -\mathbf{n}(s',t') = \left(\frac{2s}{1+s^2+t^2}, \frac{2t}{1+s^2+t^2}, \frac{1-s^2-t^2}{1+s^2+t^2}\right) = (n_1,n_2,n_3). \quad (3.5)$$

This further implies that the function h satisfies $h(s,t) = h(s',t')$. Putting things together we see that $T(s,t)$ and

$$T(s',t'): -\frac{1-s^2-t^2}{1+s^2+t^2}h(s',t') - (n_1x+n_2y+n_3z) = 0$$

describe the same carrier plane but have different orientations according to oppositely pointing unit normals $\mathbf{n}(s,t) = -\mathbf{n}(s',t')$. Finally we realize that the Blaschke image $\beta(F) = (n_1,n_2,n_3,n_0)$ of an LN surface has the special property that n_0 is a rational function over the unit sphere S^2 satisfying $n_0(\mathbf{x}) = -n_0(-\mathbf{x})$, if $\mathbf{x}$ and $-\mathbf{x}$ are antipodal points in S^2.

COROLLARY 3.2. *Let F be an LN surface whose tangent planes T are given by* (3.2). *There exists a reparametrization* (3.3) *which shows the rational offset property of LN surfaces. The Blaschke image $\beta(F) = (\mathbf{n}, n_0)$ is a function over S^2 which satisfies $n_0(\mathbf{x}) = -n_0(-\mathbf{x})$ for $\mathbf{x} \in S^2$.*

4. Laguerre geometry approach. Thinking about surface offsets as wave fronts in different time moments naturally leads to the notion of 4-dimensional Minkowski space. This section explains duality between the Gaussian sphere and the Blaschke cylinder in terms of projective Minkowski space and its dual. Relations with the three main models of the classical Laguerre geometry are established.

4.1. Gaussian sphere in projective Minkowski space. The classical 4-dimensional *Minkowski* space $\mathcal{M}$ is an affine space $\mathbb{R}^4$ with a *Minkowski* scalar product defined for every pair of vectors $\mathbf{v}$ and $\mathbf{w}$ by

$$\langle \mathbf{v}, \mathbf{w} \rangle = v_1w_1 + v_2w_2 + v_3w_3 - v_4w_4. \quad (4.1)$$

A vector $\mathbf{v}$ (or a line with a direction $\mathbf{v}$) is called isotropic (light-like) if $\langle \mathbf{v}, \mathbf{v} \rangle = 0$. The Euclidean space $\mathbb{R}^3$ will be embedded as hyperplane $x_4 = 0$ in $\mathcal{M}$. Restricting the Minkowski scalar product (4.1) to $\mathbb{R}^3$ gives the Euclidean scalar product.

The *projective Minkowski space* $\mathcal{M}P$ is a projective closure of $\mathcal{M}$ with the infinite hyperplane ω: $x_0 = 0$ containing the *absolute quadric* Ω: $x_1^2 + x_2^2 + x_3^2 - x_4^2 = 0$. Points and vectors of the affine Minkowski space $\mathcal{M}$ will be treated differently. Points $(x_1, \dots, x_4) \in \mathcal{M}$ will be identified with points $[1, x_1, \dots, x_4]$ in the affine part of $\mathcal{M}P$, and vectors $(x_1, \dots, x_4) \in \mathcal{M}$ will be used to represent points $[0, x_1, \dots, x_4]$ in ω. Then the equation of $\Omega \subset \omega$ has a compact form $\langle \mathbf{v}, \mathbf{v} \rangle = 0$, for vectors $\mathbf{v}$ in $\mathcal{M}$. Futhermore, it will be convenient to identify the absolute quadric Ω with the *Gaussian sphere* S^2 in $\mathbb{R}^3$ using the following bijective correspondence:

$$S^2 \to \Omega, \quad \mathbf{n} \mapsto \mathbf{n}^+ = (\mathbf{n}, 1). \tag{4.2}$$

An oriented surface F° in $\mathbb{R}^3$ is a surface F and a field of unit normals $\mathbf{n} : \mathbf{x} \mapsto \mathbf{n}(\mathbf{x})$, i.e. the classical Gaussian map $\mathbf{n} : F \to S^2$. By the identification (4.2) this is a map $\mathbf{n}^+ : F \to \Omega$. Define an *isotropic hypersurface* $\Gamma(F^\circ) \subset \mathcal{M}P$ as a union of isotropic lines connecting all points $\mathbf{x} \in F$ with $\mathbf{n}^+(\mathbf{x}) \in \Omega$. The affine part of any such line can be parametrized by $\mathbf{x} + t\mathbf{n}^+(\mathbf{x})$, $t \in \mathbb{R}$. Hence the orthogonal projection of any hyperplane section $\Gamma(F^\circ) \cap \{x_4 = d\}$ to $\mathbb{R}^3$ is exactly the offset F_d° of F° at the signed distance d.

The isotropic hypersurface $\Gamma(S^\circ)$ of an oriented sphere S°: $(\mathbf{x} - \mathbf{m})^2 = r^2$ in $\mathbb{R}^3$ is the union of all isotropic lines intersecting at the point $\mathbf{s} = (\mathbf{m}, \pm r) \in \mathcal{M}$, where $\mathbf{m}$ is the center and $|r|$ is the radius of S. It is easy to check that the outward pointing normals correspond to negative radius (in other sources of Laguerre geometry this choice might be opposite). Therefore, $\Gamma(S^\circ)$ coincides with the *isotropic cone* $\Gamma(\mathbf{s})$ with vertex $\mathbf{s}$ defined by the equation

$$\Gamma(\mathbf{s}) : \langle \mathbf{x} - \mathbf{s}, \mathbf{x} - \mathbf{s} \rangle = 0. \tag{4.3}$$

Laguerre transformations of $\mathcal{M}P$ are projective transformations that preserve Ω. Lines in $\mathcal{M}P$ are called space-like, isotropic, or time-like depending whether their ideal points are outside of Ω, at Ω or is inside the quadric Ω. Similarly planes and hyperplanes in $\mathcal{M}P$ are classified as space-like, isotropic, or time-like if they do not intersect Ω, are tangent to Ω or intersect Ω in more than one point, respectively.

For an oriented plane $E^\circ : e_0 + \mathbf{e}^T \cdot \mathbf{x} = 0$ in $\mathbb{R}^3$ with normal vector $\mathbf{e}$, $\|\mathbf{e}\| = 1$, its isotropic hypersurface $E^+ = \Gamma(E^\circ)$ is the hyperplane

$$E^+ : e_0 + \langle \mathbf{e}^+, \mathbf{x} \rangle = 0, \ \mathbf{x} \in \mathcal{M}. \tag{4.4}$$

Let $-E^\circ : -e_0 - \mathbf{e}^T \cdot \mathbf{x} = 0$ be the same plane E with the opposite orientation. Then $E^- = \Gamma(-E^\circ)$ is different from E^+. Both E^+ and E^- are the unique isotropic hyperplanes that contain E and are tangent to Ω.

4.2. Dual projective Minkowski space and the Blaschke cylinder. Let $\mathcal{M}P^*$ be the space dual to the projective Minkowski space $\mathcal{M}P$. Points in $\mathcal{M}P^*$ are hyperplanes $H \subset \mathcal{M}P$. The set of oriented planes in $\mathbb{R}^3$ is in 1–1 correspondence $E^\circ \mapsto E^+$ (4.4) with the set of hyperplanes in $\mathcal{M}P$ that are tangent to Ω. The latter set by duality

$$\Gamma(E^\circ) : e_0 + \langle \mathbf{e}^+, \mathbf{x} \rangle = 0 \mapsto [1, \mathbf{e}, e_0] = \beta(E^\circ) \tag{4.5}$$

defines the dual quadric $\Omega^* \subset \mathcal{M}P^*$ with the equation $y_0^2 = y_1^2 + y_2^3 + y_3^2$. The affine part $y_0 \neq 0$ of Ω^* is exactly the Blaschke cylinder defined in Section 2.2 by the equation (2.7). Note that Ω^* has just one additional real point $[0, 0, 0, 1]$ at infinity. Therefore, it is natural to call the dual Gaussian sphere Ω^* the *Blaschke cylinder* and denote it by the same letter $\mathcal{B}$.

An oriented surface F° considered as family of oriented tangent planes, defines a surface $\beta(F^\circ) \subset \mathcal{B}$, which is called the *Blaschke image* of F°. Going back to the point representation, one can check that the dual of the Blaschke image $\beta(F^\circ)^*$ is an isotropic hypersurface $\Gamma(F^\circ)$. This means that $\Gamma(F^\circ)$ can be calculated as envelope of all isotropic hyperplanes H, $H \in \beta(F^\circ)^*$.

For any surface Φ or curve in $\mathcal{M}$, define the *isotropic hypersurface* $\Gamma(\Phi) = (\Phi^* \cap \mathcal{B})^*$. This is the envelope of all isotropic hyperplanes tangent to both Φ and Ω. $\Gamma(\Phi)$ can be calculated as an envelope of all isotropic cones $\Gamma(\mathbf{x})$, $\mathbf{x} \in \Phi$, as well. The *cyclographic image* of $\Phi \subset \mathcal{M}$ in $\mathbb{R}^3$ is defined as intersection $\gamma(\Phi) = \Gamma(\Phi) \cap \mathbb{R}^3$.

REMARK 4.1. In general $\gamma(\Phi)$ is an oriented surface of two sheets. Indeed, any of its tangent planes inherits orientation from the unique isotropic tangent hyperplane of $\Gamma(\Phi)$ at the same point. There are exactly two tangent hyperplanes at any point of Φ (it is a double surface of $\Gamma(\Phi)$). For example, if F_1° and F_2° are two oriented surfaces in $\mathbb{R}^3$ then $\Phi = \Gamma(F_1^\circ) \cap \Gamma(F_2^\circ)$ is a surface in $\mathcal{M}$, and $\gamma(\Phi) = F_1^\circ \cup F_2^\circ$. The case of two cylinders will be considered in Example 6. In a recent paper [17] sufficient conditions are derived when a rational parametrization of Φ generates PN parametrizations on F_1° and F_2°.

4.3. Three models of Laguerre geometry. The classical Laguerre geometry has three main models that are described in the following table.

	Euclidean model	Cyclographic model	Blaschke model
Ambient space	Euclidean space $\mathbb{R}^3$	Minkowski space $\mathcal{M}$	Blaschke cylinder $\mathcal{B}$
Basic elements	oriented planes E° oriented spheres S°	isotropic hyperplanes points	points hyperplane sections
Basic relations	oriented contact	incidence	incidence

The correspondence between the Euclidean and the cyclographic model is defined by the maps $E^\circ \mapsto E^+$ and $S^\circ \mapsto \Gamma(S^\circ) = \Gamma(\mathbf{x}), \mathbf{x} \in \mathcal{M}$.

The *cyclographic mapping* $E^+ \mapsto \gamma(E^+)$ and $\mathbf{x} \mapsto \gamma(\mathbf{x})$ establishes the inverse correspondence. Thereby $\gamma(E^+) = E^\circ$ denotes the oriented plane $E^\circ \in \mathbb{R}^3$ corresponding to $E^+ \in \mathcal{M}$ and $\gamma(\mathbf{x})$ denotes the oriented sphere in $\mathbb{R}^3$ corresponding to the point $\mathbf{x} \in \mathcal{M}$. For the computation of the cyclographic image $\gamma(\Phi)$ of a surface $\Phi \in \mathcal{M}$ see Section 4.4.

The *Blaschke model* is focusing on the dual point of view. It is related to the cyclographic model via duality. Laguerre transformations appear in the cyclographic model as special affine transformations of $\mathcal{M}$

$$L(\mathbf{x}) = \lambda A \cdot \mathbf{x} + \mathbf{a}, \text{ with } A^T \cdot I_c \cdot A = I_c,\ I_c = \operatorname{diag}(1,1,1,-1), \lambda \in \mathbb{R} \quad (4.6)$$

where $\mathbf{x} \mapsto A\mathbf{x}$ is a *Lorentz transformation*, i.e. a linear transformation preserving the Minkowski scalar product (4.1). Laguerre transformations in the Blaschke model are defined by duality.

Laguerre transformations in the Euclidean model can be defined indirectly as follows. For an oriented surface $F^\circ \subset \mathbb{R}^3$, its Laguerre transformation $L(F^\circ)$ is computed by the formula $L(F^\circ) = L(\Gamma(F^\circ)) \cap \mathbb{R}^3$. For example, let L be the translation in x_4-direction by $-d$: $x_4 \mapsto x_4 - d$. Then $L(F^\circ) = F^\circ_d$ is the offset surface of F° at distance d.

4.4. Cyclographic images of parametric curves and surfaces in $\mathcal{M}$. A two-parameter family of spheres $S(u,v) : (\mathbf{x} - \mathbf{m}(u,v))^2 = r(u,v)^2$ with centers $\mathbf{m}(u,v)$ and radii $r(u,v)$ corresponds to a parametrized surface $\Phi : \mathbf{f}(u,v) = (\mathbf{m}, r)(u,v)$ in $\mathcal{M}$. Let $F(u,v) = \Gamma(\mathbf{f}(u,v))$ be the corresponding two-parameter family of isotropic cones with vertices $\mathbf{f}(u,v)$. Then $\Gamma(\Phi)$ is the envelope of this family, which can be computed as solution of

$$\begin{aligned} F &: \langle \mathbf{x} - \mathbf{f}, \mathbf{x} - \mathbf{f} \rangle = 0, \\ F_u &: \langle \mathbf{x} - \mathbf{f}, \mathbf{f}_u \rangle = 0, \\ F_v &: \langle \mathbf{x} - \mathbf{f}, \mathbf{f}_v \rangle = 0, \end{aligned} \quad (4.7)$$

where F_u and F_v denote the partial derivatives of F with respect to u and v. The solution of (4.7) consists of all isotropic lines that are orthogonal to Φ in the Minkowski sense (4.1).

Comparing systems (4.7) and (2.5) we recognize that the cyclographic image $\gamma(\Phi) = \Gamma(\Phi) \cap \mathbb{R}^3$ of the parametrized surface $\Phi : \mathbf{f}(u,v) = (\mathbf{m}, r)(u,v)$ is the envelope of the two-parameter system of spheres $S(u,v)$ in $\mathbb{R}^3$. We note that only the points $\mathbf{f}$ of Φ whose tangent planes T_f spanned by $\mathbf{f}_u$ and $\mathbf{f}_v$ are space-like or isotropic, will contribute to the real part of the cyclographic image $\gamma(\Phi)$.

A one-parameter family of spheres $S(t) : (\mathbf{x} - \mathbf{m}(t))^2 = r(t)^2$ corresponds to a curve $\mathbf{s}(t) = (\mathbf{m}, r)(t)$ in $\mathcal{M}$. By similar calculations the isotropic hypersurface $\Gamma(\mathbf{s})$ consists of all isotropic lines that intersect the curve orthogonally. The cyclographic image $\gamma(\mathbf{s}) = \Gamma(\mathbf{s}) \cap \mathbb{R}^3$ is the envelope of the family of spheres $S(t)$ and is called *canal surface*. It is real exactly if tangent vectors $\dot{\mathbf{s}}$ are space-like or isotropic

$$\langle \dot{\mathbf{s}}(t), \dot{\mathbf{s}}(t) \rangle = ||\dot{\mathbf{m}}(t)||^2 - \dot{r}(t)^2 \geq 0. \quad (4.8)$$

5. Universal rational parametrizations of the sphere and the Blaschke cylinder. Dietz, Hoschek and Jüttler [6] noticed that Bézier curves and surface patches on the unit sphere S^2 can be represented uniformly by introducing the *generalized stereographic projection* $\delta : \mathbb{R}^4 \to S^2$

$$\delta(a,b,c,d) = (a^2+b^2+c^2+d^2, 2ac+2bd, 2bc-2ad, a^2+b^2-c^2-d^2). \quad (5.1)$$

In complex notations $z = a + b\mathrm{i}$, $w = c + d\mathrm{i}$ this construction has the following form [18]

$$P_S(z,w) = (|z|^2 + |w|^2, 2\mathrm{Re}(z\bar{w}), 2\mathrm{Im}(z\bar{w}), |z|^2 - |w|^2), \quad (5.2)$$

and is called a *universal rational parametrization* of S^2 (see [4, 19] for details). Since $P_S : \mathbb{C}^2 \to S^2$ is homogeneous $P_S(\lambda z, \lambda w) = |\lambda|^2 P_S(z,w)$, $\lambda \in \mathbb{C}$, P_S defines also a map from a complex projective line $\mathbb{C}P^1$ to S^2, which is essentially the Riemann sphere construction.

The universal property of P_S is formulated in the following theorem. Here we call a finite collection of polynomials $(f_0, f_1, \ldots)$ *irreducible* if $\gcd(f_0, f_1, \ldots) = 1$.

THEOREM 5.1. *Any irreducible solution* $\mathbf{f} = (f_0, \ldots, f_3) \in \mathbb{R}[t_1, \ldots, t_k]^4$ *of the unit sphere equation* $f_0^2 = f_1^2 + f_2^2 + f_3^2$ *has the form* $\mathbf{f} = P_S(\mathbf{F})$ *with an irreducible* $\mathbf{F} = (z,w) \in \mathbb{C}[t_1, \ldots, t_k]^2$*, which is determined uniquely up to a complex constant multiplier* λ*,* $|\lambda| = 1$.

We call $\mathbf{F} = (z,w)$ a *lifting* of $\mathbf{f}$ and denote it by $\widetilde{\mathbf{f}} = \mathbf{F}$. The lifting can be calculated using a simple formula [20]

$$\widetilde{\mathbf{f}} = \big(h(f_0 + f_3)/(f_1 - f_2\mathrm{i}), h\big), \quad h = \gcd(f_1 - f_2\mathrm{i}, f_0 - f_3). \quad (5.3)$$

The formula (5.3) enables the lifting of rational Bézier curves of degree $2k$ and tensor product patches of degree $(2k, 2l)$ on S^2 to the corresponding polynomial curves of degree k and surfaces of bi-degree (k,l) in $\mathbb{C}^2$, respectively. This universal rational parametrization technique was used to find Bézier patches on S^2 of minimal degree with given boundary curves, see [19]. Theorem 5.1 can be applied for polynomials of arbitrary number of variables.

EXAMPLE 2. *Consider a parametrization* $\mathbf{f}$ *which is the opposite on* S^2 *to* $P_S(z,w)$*:*

$$\mathbf{f} = (-|z|^2 - |w|^2, 2\mathrm{Re}(z\bar{w}), 2\mathrm{Im}(z\bar{w}), |z|^2 - |w|^2).$$

Let us calculate a lifting: $h = \gcd(2\bar{z}w, -2|z|^2) = \bar{z}$*, and* $\widetilde{\mathbf{f}} = (\bar{z}(-2|w|^2)/(2\bar{z}w), \bar{z}) = (-\bar{w}, \bar{z})$*. Therefore, a point* $P_S(-\bar{w}, \bar{z})$ *is the opposite of* $P_S(z,w)$*.*

Projectively the Blaschke cylinder $\mathcal{B}$ is a cone over a sphere, since its equation is $y_0^2 = y_1^2 + y_2^2 + y_3^3$ (y_4 is arbitrary) in $\mathcal{M}P$. In complex setting this equation can be transformed to the binomial one $(y_0 - y_3)(y_0 + y_3) =$

$(y_1 - \mathrm{i}y_2)(y_1 + \mathrm{i}y_2)$. Therefore $\mathcal{B}$ is a real part of a toric variety (see e.g. [49]). According to a general theory [4], $\mathcal{B}$ has the universal rational parametrization in the slightly more complicated form:

$$P_B(z,w,f,g) = \big(g(|z|^2+|w|^2), 2g\mathrm{Re}(z\bar{w}), 2g\mathrm{Im}(z\bar{w}), g(|z|^2-|w|^2), f\big). \quad (5.4)$$

The map $P_B : \mathbb{C}^2 \times \mathbb{R}^2 \to \mathbb{R}^5$ is homogeneous, $P_B((\lambda,\rho) * (z,w,f,g)) = |\lambda|^2 \rho P_B(z,w,f,g)$, with respect to the following multiplication:

$$(\lambda,\rho) * (z,w,f,g) = (\lambda z, \lambda w, |\lambda|^2 \rho f, \rho g), \quad \lambda \in \mathbb{C}, \rho \in \mathbb{R}. \quad (5.5)$$

THEOREM 5.2. *Any irreducible solution* $\mathbf{h} = (h_0, \dots, h_4) \in \mathbb{R}[t_1, \dots, t_k]^4$ *of the Blaschke cylinder equation* $h_0^2 = h_1^2 + h_2^2 + h_3^2$ *has the form* $\mathbf{h} = P_B(\mathbf{H})$*, where* $\mathbf{H} = (z,w,f,g) \in \mathbb{C}[t_1, \dots, t_k]^2 \times \mathbb{R}[t_1, \dots, t_k]^2$ *and the pairs* (z,w) *and* (f,g) *are irreducible.* $\mathbf{H}$ *is determined uniquely up to multiplication by* (λ,ρ) *defined in* (5.5)*, with* $|\lambda|^2\rho = 1$.

EXAMPLE 3. *Consider the particular parametrization of the Blaschke cylinder* $\iota : I^3 \to \mathcal{B}$*:*

$$\iota : I^3 = \mathbb{R}^3 \cup \mathbb{R} \to B, \quad (u,v,f) \mapsto P_B(u + v\mathrm{i}, 1, f, 1), \ f \mapsto P_B(1,0,f,1),$$

which is called the isotropic *model of Laguerre geometry. The composition with the Blaschke map* $\Lambda = \iota^{-1} \circ \beta$ *describes the change from the Euclidean model to the isotropic model. It will be useful for modeling applications in Section 7.1.*

6. Special cases of PN surfaces. Important examples of PN surfaces are generated as envelopes of rational one parameter families of simplest primitive shapes: planes, spheres or circular cones (cylinders).

Envelopes of planes are exactly developable surfaces. Developable PN surfaces were already characterized by Theorem 2.2 as rational curves in the Blaschke cylinder (see Section 2.3). The other classes of PN surfaces are now discussed in the subsequent sections.

6.1. Canal surfaces. A canal surface is the envelope of one-parameter family of spheres in $\mathbb{R}^3$ defined by a *spine curve* $\mathbf{m}(t)$ and a *radius function* $r(t)$. In 1995 Lü [27] proved the surprising result: *A canal surface defined by a rational spine curve and a rational radius function is rational.* See also [28] for the details of the proof.

Later this result was proved by different methods: geometric approach [34], Clifford algebra formalism [2, 3] and a universal rational parametrization of a sphere [20]. Here we describe the approach in [34, 20] which gives bounds of rational parametrization degree.

As it was explained in Section 4.4, a canal surface is the cyclographic image $\gamma(\mathbf{s})$ of a curve $\mathbf{s}(t) = (\mathbf{m}, r)(t)$ in $\mathcal{M}$. It is real exactly if $\langle \dot{\mathbf{s}}(t), \dot{\mathbf{s}}(t) \rangle \geq 0$ (see (4.8)), and its isotropic hypersurface $\Gamma(\mathbf{s})$ consists of all isotropic lines that intersect the curve orthogonally. Therefore, we can look for a

parametrization of $\Gamma(\mathbf{s})$ in the form $\mathbf{F}(s,t,\lambda) = \mathbf{s}(t) + \lambda\mathbf{n}^+(s,t)$, where $\langle\dot{\mathbf{s}}(t), \mathbf{n}^+(s,t)\rangle = 0$. The latter condition means that isoparametric curves $\mathbf{a}_t(s) = \mathbf{n}(s,t)$ of a Gauss map $\mathbf{n}(s,t)$ define a family of circles as planar sections of the unit sphere S^2,

$$\langle\dot{\mathbf{s}}(t), \mathbf{x}^+\rangle = 0, \quad \mathbf{x}^+ = (\mathbf{x}, 1),\ \mathbf{x} \in \mathbb{R}^3. \tag{6.1}$$

If such a rational Gauss map $\mathbf{n}(s,t)$ exists then a rational parametrization $\mathbf{f}(s,t)$ of the canal surface $\gamma(\mathbf{s}) = \Gamma(\mathbf{s}) \cap \mathbb{R}^3$ can be calculated by substitution $\mathbf{f}(s,t) = \mathbf{F}(s,t,-s_4(t))$. Equivalently, $\mathbf{f}(s,t)$ can be expressed in terms of the spine curve $\mathbf{m}(t)$ and the radius function $r(t)$

$$\mathbf{f}(s,t) = \mathbf{m}(t) - r(t)\mathbf{n}(s,t). \tag{6.2}$$

Consider the slightly more general case. Let $\Pi_t : \langle\mathbf{v}(t), \mathbf{x}^+\rangle = 0$ be a family of planes with polynomial coefficients $\mathbf{v}(t)$, where $D(t) = \langle\mathbf{v}(t), \mathbf{v}(t)\rangle \geq 0$, $D(t) \not\equiv 0$. Then the polynomial $D(t)$ can be factorized

$$D(t) = \prod_i (t - z_i)^{p_i}(t - \bar{z}_i)^{p_i}\rho(t)^2, \quad z_i \in \mathbb{C} \setminus \mathbb{R}, \quad \rho(t) \in \mathbb{R}[t]. \tag{6.3}$$

THEOREM 6.1 ([20]). *There exists a rational parametrization $\mathbf{n}(s,t)$ of the unit sphere S^2 of bidegree $(2,n)$, such that all isoparametric curves $\mathbf{a}_t(u) = \mathbf{n}(s,t)$ are plane sections $S^2 \cap \Pi_t$. There is a constructive method for finding such parametrization $\mathbf{n}(s,t)$ with*

$$n = 2\max(\lceil m/2\rceil, m - \sum_i \lceil p_i/2\rceil), \quad m = \deg\mathbf{v}(t), \tag{6.4}$$

where p_i are multiplicities of complex roots of the polynomial $D(t)$ defined in (6.3).

The proof is based on the universal rational parametrization of the sphere (5.2), where P_S is treated as a map $\mathbb{R}P^3 \to S^2$. For any representation of $D(t)$ as a sum of squares there is a *minimal* solution $\mathbf{n}(s,t) = P_S(\mathbf{r}(s,t))$, which is an image of a certain ruled surface $\mathbf{r}(s,t)$ of implicit degree m in $\mathbb{R}P^3$. Minimal directrices of $\mathbf{r}(s,t)$ with complimentary degrees $m_0 + m_1 = m$ can be found using μ-basis methods [1]. The goal is to find the minimal $n = 2\max(m_0, m_1)$ in the list of all minimal solutions, which is achieved by a constructive procedure up till the value (6.4).

REMARK 6.1. The related problem of finding a decomposition of a real polynomial as a sum of two squares over $\mathbb{Q}$ was considered in [25]. It was proved that the problem is equivalent to partial factorization of of the polynomial, and a decomposition algorithm was presented in case the solution is defined over $\mathbb{Q}$.

Theorem 6.1 can be applied to a family of planes (6.1) with $\mathbf{v}(t) = d(t)\dot{\mathbf{c}}(t)$, when $d(t)$ is a common denominator of all rational $\dot{c}_i(t)$, $i =$

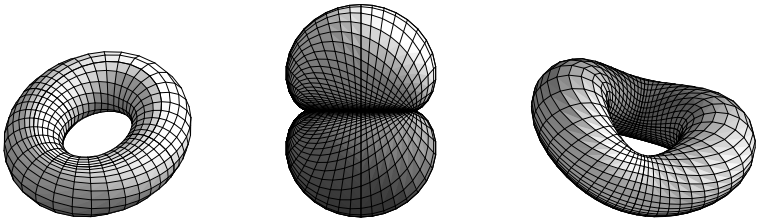


FIG. 2. *Examples of canal surfaces of bidegree* $(2,4)$, $(2,6)$, $(2,10)$.

$1,\dots,4$. If $\mathbf{c}(t)$ is a rational curve of degree k then in general $\deg \mathbf{v}(t) = 2k-2$. Using the inequality $\deg_t \mathbf{f}(s,t) \le \deg \mathbf{c}(t) + \deg_t \mathbf{n}(s,t)$ (see (6.2)) one can derive the following degree bounds.

COROLLARY 6.1 ([20]). *A canal surface* $\gamma(\mathbf{c})$ *defined by a rational curve* $\mathbf{c}(t)$ *in* $\mathbb{R}^4$ *of degree* k *admits a rational parametrization* $\mathbf{f}(s,t)$ *of bidegree* $(2,n)$, *where*

(i) $n = 3k-2$ *if* $D(t) > 0$ *(all roots are complex),*
(ii) $n = 5k-6$ *if* $D(t)$ *has at least one complex root.*
(iii) $n = 5k-4$ *if* $D(t)$ *has all real roots.*

The case (i) of this corollary gives in general the minimal possible degree, and the case (ii) was proved in [34]. The following example shows that and the case (iii) cannot be improved.

EXAMPLE 4. Consider a canal surface $\gamma(\mathbf{c})$ (see Fig. 2 middle) defined by the curve

$$\mathbf{c}(t) = \Big(0, \frac{1-t^2}{2(1+t^2)}, \frac{t}{1+t^2}, -\frac{t}{1+t^2}\Big).$$

Then $\mathbf{v}(t) = (1+t^2)^2\dot{\mathbf{c}}(t) = (0,-2,1-t^2,-1+t^2)$ and $D(t) = 4t^2$ has no complex roots. Therefore there is only one factorization of $D(t)$ that defines a unique minimal parametrization $\mathbf{n}(t,u) = P_S((1-s)X + sY)$, with $X = (\mathrm{i}(t^2-1), 2t)$, $Y = (1,0)$, of degree $(2,4)$. The canal surface $\gamma(\mathbf{c})$ is parametrized by $\mathbf{f}(s,t)$ (see (6.2)) of bidegree $(2,6)$, which is minimal possible.

6.2. Rational ruled surfaces in $\mathcal{M}$. Rational ruled surfaces in $\mathbb{R}^3$ are PN surfaces if they are non-developable. This result was proved in [35, 31] and generalized to any rational ruled surface Ψ in $\mathcal{M}$ in the sense that its cyclographic image $\gamma(\Psi)$ in $\mathbb{R}^3$ is rational.

The Blaschke model will be most convenient for understanding this result. Let a rational ruled surface Ψ in $\mathcal{M}$ be defined by two directrices $\mathbf{c}(t)$ and $\mathbf{d}(t)$. The key idea is to construct a Gaussian map $\mathbf{n}(s,t)$ which for any fixed t defines normals along a common touching curve between the envelope surface $\gamma(\Psi)$ and a cone of revolution defined by a line going through points $\mathbf{c}(t)$ and $\mathbf{d}(t)$. This is equivalent to the condition

$$\langle \mathbf{n}^+(s,t), \mathbf{d}(t) - \mathbf{c}(t) \rangle = 0, \tag{6.5}$$

which forces isoparametric curves of $\mathbf{n}(s,t)$ to be the prescribed family of circles. Therefore, such $\mathbf{n}(s,t)$ can be generated using Theorem 6.1. Then a support function $h(s,t)$ is computed from the equation

$$h(s,t) + \langle \mathbf{n}^+(s,t), \mathbf{c}(t) \rangle = 0. \tag{6.6}$$

Finally it remains to go back from the Blaschke image to $\gamma(\Psi)$, i.e. to compute the envelope of tangent planes $h(s,t) + \langle \mathbf{n}(s,t), \mathbf{x} \rangle = 0$ in $\mathbb{R}^3$.

EXAMPLE 5. Let Ψ be a hyperbolic paraboloid $x_3 = x_1 x_2$ in $\mathbb{R}^3$ with two directrices $\mathbf{c}(t) = (0,t,0,0)$ and $\mathbf{d}(t) = (1,t,t,0)$. The Gaussian map can be calculated as

$$\begin{aligned} \mathbf{n}(s,t) &= P_S(-st + \mathrm{i}(1-s), 1 - 3s - \mathrm{i}st) \\ &= \frac{(-4st + 8s^2 t, -2s^2t^2 + 2 - 8s + 6s^2, 4s - 8s^2)}{(1 - 4s + 5s^2 + s^2t^2)}. \end{aligned} \tag{6.7}$$

One can check directly that the condition (6.5) is fulfilled. Then the Blaschke image is derived using (6.6), and finally a point representation of any offset of Ψ is generated of bidegree $(4,5)$. Compare with a parametrization of bidegree $(6,6)$ that is generated by treating Ψ as LN surface (see Section 3).

6.3. Characterization of PN surfaces of low parametrization degree. For a PN surface F parametrized by $\mathbf{f}(s,t)$ consider the parametrization of its isotropic hypersurface $\Gamma(F)$ in the form $\mathbf{F}(s,t,u) = \mathbf{f}(s,t) + u\mathbf{n}^+(s,t)$ (see Section 6.1), and a family Φ_t of its isoparametric ruled subsurfaces $\Phi_t(s,u) = \mathbf{F}(s,t,u)$. Define the *PN degree* of the PN parametrization of F with respect to s as implicit degree of Φ_t. Then any general Laguerre transform of $\mathbf{f}(s,t)$ (see Section 4.3) will have the same PN degree in s.

It was shown in previous Sections 6.1 and 6.2 that the simplest PN surfaces admit parametrizations $\mathbf{f}(s,t)$ with $\deg_s(\mathbf{f}) \le 4$. Now we are going to show that they can be almost characterized by this degree.

THEOREM 6.2. *If a PN surface F parametrized by $\mathbf{f}(s,t)$ has PN degree $m \le 4$ in s then F is:*

(i) *a developable PN surface if $m = 1$;*
(ii) *a rational canal surface if $m = 2$;*
(iii) *an envelope of a rational family of circular cones if $m = 3$;*
(iv) *an envelope of a rational family of circular cones or Dupin cyclides if $m = 4$.*

Proof. If $m = 1$ then the surface Φ_t is a plane and its projection to $\mathbb{R}^3$ is also a plane containing normals $\mathbf{n}(s,t)$ along the line $\Phi_t(s,0)$. Hence $\mathbf{n}(s,t)$ is constant for any fixed t, and F is developable.

In the case $m = 2$ the surface Φ_t is a quadric that spans a time-like hyperplane in $\mathcal{M}$, that can be moved to a standard position $x_3 = 0$ using an appropriate Laguerre transformation. Then all normals along the conic $\Phi_t(s, 0)$ will be in the same plane. Thus these conics are circles and Φ_t are cones with vertices $\mathbf{c}(t)$ that trace a curve in $\mathcal{M}$. Then F is a canal surface $\gamma(\mathbf{c})$.

In the case $m = 3$ the surface Φ_t is a ruled cubic, which cannot be a cone (since otherwise its vertex belongs to Ω and F is developable). Then there is exactly one linear directrix $\mathbf{l}_t$ in Φ_t for every parameter t, i.e. they define a rational family of circular cones $\gamma(\mathbf{l}_t)$ with the envelope F.

In the last case $m = 4$ the surface Φ_t is a ruled quartic. By the same arguments as above this cannot be a cone. Then Φ_t has a family of conics $\mathbf{c}_v$ as directrices. Since all canal surfaces $\gamma(\mathbf{c}_v)$ are touching along a common quartic curve $\Phi_t(s, 0)$, conics $\mathbf{c}_v$ have only space-like tangents. Such canal surfaces have been studied in [24], where it has been proved that there exists a circular cone in the family $\gamma(\mathbf{c}_v)$, except in the Dupin cyclide case (see [24], Corollary 2). Therefore F is an envelope of these circular cones. □

Examples with $m = 3, 4$ are provided by branching blend surfaces of bidegree $(3, 6)$ and $(4, 8)$ in Section 7.4 (see also [21]). The latter case corresponds to a family of Dupin cyclides.

6.4. Offsets of regular quadric surfaces in $\mathbb{R}^3$. Regular quadrics are one of the simplest surfaces in $\mathbb{R}^3$. Nevertheless it is not obvious that their offsets admit rational parametrizations. Investigating conics in the plane it is quite clear that the offset curves of ellipses and hyperbolas are non-rational whereas the offsets of circles and parabolas are rational curves. Lü [30] has been the first who proved that the offsets of all regular quadrics admit rational parametrizations. For the paraboloids this is not difficult, for ellipsoids and hyperboloids this is quite involved.

The existence of rational parametrizations of the offsets of regular quadrics can be shown as follows. Let Φ be a two-dimensional quadric in $\mathcal{M}$, then Φ is contained in a hyperplane of $\mathcal{M}$. We are studying the isotropic hypersurface $\Gamma(\Phi)$ corresponding to Φ. The intersection of $\Gamma(\Phi)$ with $x_4 = 0$ is the cyclographic image $\gamma(\Phi)$, the envelope of the two-parameter family of oriented spheres corresponding to Φ. If Φ is contained in a hyperplane $x_4 = d$, the envelope $\gamma(\Phi)$ is the offset surface of Φ.

The isotropic hypersurface $\Gamma(\Phi)$ is the envelope of common tangent hyperplanes of the pencil of quadrics $\lambda\Phi + \mu\Omega$ in $\mathcal{M}P$. The quadrics Φ and Ω are considered as sets of tangent hyperplanes, which implies that they are singular hypersurfaces in this pencil.

The intersection surface of two hyperquadrics in $\mathcal{M}P$ is a rational quartic del Pezzo surface. This del Pezzo surface is dual to the isotropic hypersurface $\Gamma(\Phi)$. This implies that the two parameter family of tangent hyperplanes of $\Gamma(\Phi)$ can be rationally parametrized. Intersecting $\Gamma(\Phi)$

with $x_4 = 0$ gives a parametrization of $\gamma(\Phi)$ as set of tangent planes. This construction proves the following result.

PROPOSITION 6.1. *The cyclographic images $\gamma(\Phi)$ of two dimensional quadrics $\Phi \subset \mathcal{M}$ are surfaces admitting rational parametrizations.*

We demonstrate also an alternative way to construct rational PN parametrizations of the offset surfaces of quadrics in $\mathbb{R}^3$. Let Φ be a quadric surface in $\mathcal{M}$, contained in a space-like hyperplane, for instance $x_4 = 0$. We show that the pencil of quadrics $\lambda\Phi + \mu\Omega$ in $\mathcal{M}P$ contains a ruled quadric surface Ψ and $\Gamma(\Phi) = \Gamma(\Psi)$ holds. Then the cyclographic images $\gamma(\Phi)$ and $\gamma(\Psi)$ agree and rational parametrizations can be constructed as described in Section 6.2.

THEOREM 6.3. *All regular quadrics are PN surfaces.*

Let Φ be a regular quadric possessing real points. Quadrics of revolution are canal surfaces and thus the rationality of their offsets follows from Section 6.1. If Φ itself is a ruled quadric surface in $\mathcal{M}$, we may directly apply the method outlined in Section 6.2.

Otherwise let Φ be contained in the hyperplane $x_4 = 0$. The pencil of dual hyperquadrics $\lambda\Phi + \mu\Omega$ in $\mathcal{M}$ defines the isotropic hypersurface $\Gamma(\Phi)$. All singular quadrics in this pencil possess the same isotropic hypersurface $\Gamma(\Phi)$. The offset surfaces of Φ at distance d are obtained as hyperplane sections $\Gamma(\Phi) \cap \{x_4 = d\}$. We will find a real ruled quadric Ψ in all of the three cases which have to be discussed.

- Let Φ be an *ellipsoid* in $x_4 = 0$. Then Φ (with $a > b > c$) and Ψ are given by the equations

$$\Phi\colon \frac{x_1^2}{a^2} + \frac{x_2^2}{b^2} + \frac{x_3^2}{c^2} = 1, x_4 = 0, \quad \Psi\colon \frac{x_1^2}{a^2 - b^2} - \frac{x_3^2}{b^2 - c^2} + \frac{x_4^2}{b^2} = 1, x_2 = 0.$$

- Let Φ be a *two sheet hyperboloid* in $x_4 = 0$. Then Φ (with $b > c$) and Ψ are given by the equations

$$\Phi\colon \frac{x_1^2}{a^2} - \frac{x_2^2}{b^2} - \frac{x_3^2}{c^2} = 1, x_4 = 0, \quad \Psi\colon \frac{x_1^2}{a^2 + b^2} + \frac{x_3^2}{b^2 - c^2} - \frac{x_4^2}{b^2} = 1, x_2 = 0.$$

- Let Φ be an *elliptic paraboloid* in $x_4 = 0$. Then Φ (with $b > c$) and the hyperbolic paraboloid Ψ are given by the equations

$$\Phi\colon \frac{x_1^2}{a^2} + \frac{x_2^2}{b^2} - 2x_3 = 0, x_4 = 0, \quad \Psi\colon -\frac{1}{a^2 - b^2}x_2^2 + \frac{1}{a^2}x_4^2 = 2x_3 - a, x_1 = 0.$$

6.5. Quadratic triangular Bézier surfaces in $\mathcal{M}$. In Section 6.2 it has been proved that the cyclographic images $\gamma(F)$ of rational ruled surfaces F in $\mathcal{M}$ are PN surfaces. Besides this result not much has been known about rationality of cyclographic images. Recently it has been proved in [36] that any quadratic triangular Bézier surface in $\mathcal{M}$ possesses a rational envelope surface of the corresponding family of spheres. This result can

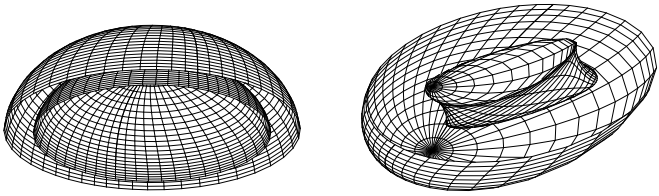


FIG. 3. *Left: Ellipsoid of revolution and outside offset. Right: General ellipsoid and inside offset.*

directly be proved starting with an appropriate parametrization $\mathbf{w}(s,t)$ of the absolute quadric Ω and solving the equations (4.7).

Let

$$\mathbf{f}(u,v) = \frac{1}{2}\mathbf{a}_1 u^2 + \mathbf{a}_2 uv + \frac{1}{2}\mathbf{a}_3 v^2 + \mathbf{a}_4 u + \mathbf{a}_5 v + \mathbf{a}_6, \text{ with } \mathbf{a}_i \in \mathbb{R}^4 \quad (6.8)$$

be a parametrization of a quadratic triangular Bézier surface F spanning $\mathbb{R}^4$. For convenience we use the monomial basis instead of the Bernstein basis for the representation of F.

In order to solve (4.7), we may start with a rational parametrization $\mathbf{w}(s,t)$ of Ω, which obviously satisfies $\langle \mathbf{w}, \mathbf{w} \rangle = 0$. A possible choice is $\mathbf{w}(s,t) = (2s, 2t, 1-s^2-t^2, 1+s^2+t^2)$. The conditions $\langle \mathbf{w}, \mathbf{f}_u \rangle = 0$ and $\langle \mathbf{w}, \mathbf{f}_v \rangle = 0$ are linear in u and v. Thus, a solution of the system of linear equations

$$\begin{pmatrix} \langle \mathbf{w}, \mathbf{a}_1 \rangle & \langle \mathbf{w}, \mathbf{a}_2 \rangle \\ \langle \mathbf{w}, \mathbf{a}_2 \rangle & \langle \mathbf{w}, \mathbf{a}_3 \rangle \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} -\langle \mathbf{w}, \mathbf{a}_4 \rangle \\ -\langle \mathbf{w}, \mathbf{a}_5 \rangle \end{pmatrix} \quad (6.9)$$

is a rational reparametrization $u = a(s,t)$, $v = b(s,t)$ for the quadratic triangular Bézier surface F. It can be proved that the determinant of the coefficient matrix of (6.9) does not vanish identically expect for quadratically parameterized planes F. The isotropic lines $i(s,t) : \mathbf{f}(s,t) + \lambda \mathbf{w}(s,t)$ are solutions of (4.7) and form a rational parametrization of the isotropic hypersurface $\Gamma(F)$ through F. The intersection $\Gamma(F) \cap \mathbb{R}^3$ is the envelope $\gamma(F)$ of the two-parameter family of spheres corresponding to F.

7. Modeling applications. Our first non-trivial modeling applications of surfaces with rational offsets are related to Dupin cyclides (see Fig. 4, left). These are special canal surfaces which are cyclographic images of Minkowski circles, i.e. conics in $\mathcal{M}$ that intersect Ω in two points (see details in [5, 23]). Dupin cyclides were proposed to be used as blending surfaces between natural quadrics by Pratt [46, 47] (see Fig. 4). For example, any two circular cones with a common inscribed sphere can be blended by a part of a Dupin cyclide bounded by two circles (Fig. 4, middle). In terms of the cyclograpic model this is a simple rounding of two intersecting space-like lines by an arc of a Minkowski circle. Similar blending is available between a circular cylinder (or cone) and a plane (Fig. 4, right).

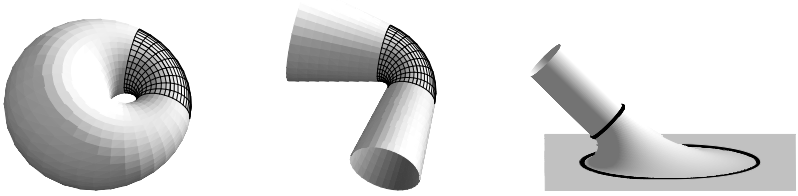


FIG. 4. *Using patches of cyclides for blending.*

More sophisticated modeling schemes with patches of Dupin cyclides bounded by circles were considered in [50]. A very special of blending between two circular cylinders using parabolic Dupin cyclides was proposed in [51].

7.1. Modeling with parabolic cyclides. The technique described in Example 1 can be used to develop a surface modeling scheme based on parabolic Dupin cyclides in [33] . Let scattered data elements $(\mathbf{a}_i, A_i)$ be given in $\mathbb{R}^3$, where $\mathbf{a}_i$ are vertices incident with the oriented planes A_i. The goal is to construct a C^1 PN surface, which interpolates the given data and which is composed of triangular patches of parabolic Dupin cyclides. The concept is the following. The data $(\mathbf{a}_i, A_i,)$ are mapped by Λ to the isotropic model I^3 (see also Example 3). The images are scattered data elements, say $(\mathbf{b}_i, B_i)$, with $\Lambda(A_i) = B_i$. The data $(\mathbf{b}_i, B_i)$ will be interpolated by a C^1 function Ψ, which is piecewise quadratic, using the method of Powell–Sabin [45]. Returning to the standard model we obtain a C^1 interpolating surface $\Lambda^{-1}(\Psi)$ composed of parabolic Dupin cyclides. We note that in general the triangular cyclide pieces are tangent to each other along cubics and not along circles. This already indicates that this method is rather different from other surface modeling schemes, using (parabolic) Dupin cyclides, as [46, 50] and others. However, smooth surfaces with vanishing Gaussian curvature along curves other than straight lines can never be modeled with parabolic Dupin cyclides.

7.2. Approximations with developable PN surfaces. Very few applications of developable PN surfaces are known. In [26] developable surfaces are modeled with pieces of circular cones. A more general method for the recognition and reconstruction of developable surfaces was proposed in [32]. The approximation problem with given data points as measurements from a developable surface and estimated tangent planes is translated to a curve fitting problem which is solved on the Blaschke cylinder. Then the constructed curve is interpreted as one-parameter family of tangent planes and their envelope is calculated.

7.3. Blending natural quadrics with canal surfaces. Canal surfaces defined by general conics in $\mathcal{M}$ can be used for blending

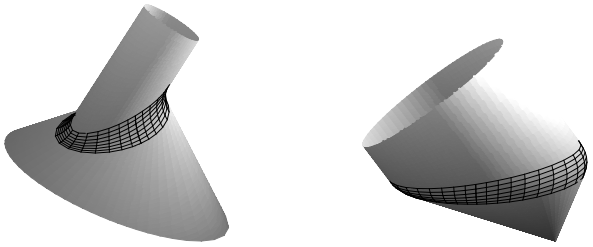


FIG. 5. *Blendings of cylinders/cones of bidegree* $(2, 4)$ *along quartic curves.*

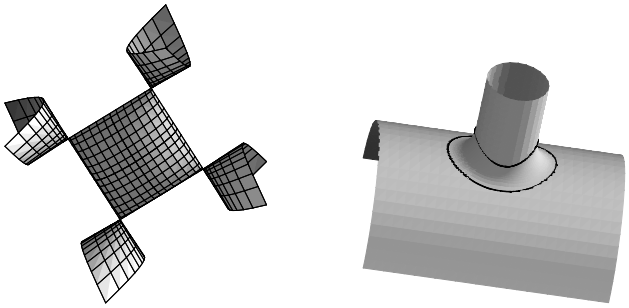


FIG. 6. *Bisector of two cylinders and their blending of bidegree* $(2, 10)$.

cones/cylinders in more general positions (Fig. 5) along quartic boundary curves as was shown in [24].

Results on rational parametrizations [34, 20] have been used in [15] to develop a theory on rational variable radius rolling ball blends between natural quadrics in arbitrary positions. Here we will consider just one illustrative example.

EXAMPLE 6. Let Q_a and Q_b be two cylinders in $\mathbb{R}^3$ defined by equations $x_1^2 + x_2^2 = r_a^2$ and $x_2^2 + x_3^2 = r_b^2$, where $0 < r_a < r_b$ (Fig. 6, right). Consider lines L_a and L_b in $\mathcal{M}$ such that their cyclographic images are given cylinders Q_a and Q_b. Then an intersection of their isotropic hypersurfaces is a quartic surface $\Phi = \Gamma(L_a) \cap \Gamma(L_b)$ in $\mathcal{M}$ which projects exactly to the bisector of the cylinders in $\mathbb{R}^3$ (Fig. 6, left).

Any curve on Φ defines a canal surface touching both cylinders, i.e. a rolling ball blend. Unfortunately a fixed radius case corresponds to a non-rational curve on Φ. Nevertheless, a certain rational quartic curve $\mathbf{s} \subset \Phi$ can be found [22]. This construction generates a canal surface of bidegree $(2, 10)$ which is minimal possible according to Corollary 6.1. It is impossible to construct such a blending with a boundary circle on the cylinder Q_a, since the corresponding curve on Φ and the associated canal surface are non-rational.

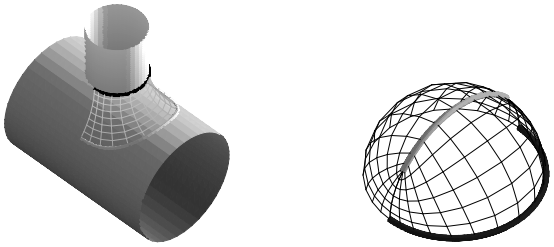


FIG. 7. *Boundary curves C_a, C_b and their images $\mathbf{n}_a$, $\mathbf{n}_b$ on the Gaussian sphere.*

7.4. Branching blend of natural quadrics using non-canal PN surfaces. The blending in Example 6 can be improved by using PN surfaces that are more general than canal surfaces. Here we sketch the construction proposed in [21].

The goal is to generate a branching blend of cylinders Q_a and Q_b defined in Example 6, which is a PN surface bounded by a circle C_a, $x_3 = h$, on the vertical cylinder Q_a and by a rational curve C_b on the upper side of the horizontal cylinder Q_b (see Fig. 7, left).

The general scheme of the proposed method consists of three steps.

Step 1: *Gaussian map.* Normals along C_a and C_b define two curves on the unit sphere: a circle $\mathbf{n}_a$ on the equator and a circular arc $\mathbf{n}_b$ on the plane section $x_1 = 0$, see Fig. 7 (right). In order to build a symmetric Gaussian map it remains to find a Bézier representation of the spherical quarter. Methods of [19] can be directly applied: a linear combination of the liftings $\widetilde{\mathbf{n}}_a$ and $\widetilde{\mathbf{n}}_b$ to $\mathbb{C}^2$ (see Section 5)

$$(z, w) = (1-s)(1-\mathrm{i}t)\widetilde{\mathbf{n}}_a + s\widetilde{\mathbf{n}}_b, \quad \widetilde{\mathbf{n}}_a = (1-\mathrm{i}t, t-\mathrm{i}),\ \widetilde{\mathbf{n}}_b = (1+t^2, 2kt). \quad (7.1)$$

with the resulting unique parametrization in a homogeneous form $\mathbf{n}(s,t) = P_S(z,w)$ of degree $(2,4)$, where homogeneous coordinates are used $\mathbf{n} = (n_0, \ldots, n_3)$. Note, that the parameter k controls the endpoints of the arc $\mathbf{n}_b$.

Step 2: *Support function.* Points $\mathbf{s}_a = (0,0,h,r_a)$ and $\mathbf{s}_b = (0,0,0,r_b)$ in Minkowski space $\mathcal{M}$ represent two spheres: touching the cylinder Q_a along C_a, and the cylinder Q_b along a circle with the normal $\mathbf{n}_b$. Their Blaschke images are constructed with the same fixed Gaussian map $\mathbf{n}(s,t)$ and represented in the universal rational parametrization form (5.4) with $g_a = g_b = 1$ and certain polynomials f_a and f_b of bidegree $(2,4)$. The formula $f(s,t) = f_a(s,t) + s^2(f_b(1,t) - f_a(1,t))$ defines a polynomial that is in C^1-contact with f_a along $s = 0$ and coincides with f_b on $s = 1$. Then the parametrization $(\mathbf{n}(s,t), f(s,t)) = P_B(z,w,f,1)$ is the dual of the blending solution.

Step 3: *Back to the point representation.* From the dual representation $e_0(s,t) + \langle \mathbf{e}(s,t), \mathbf{x}\rangle = 0$, $e_0 = f$, $\mathbf{e} = (n_1, n_2, n_3)$, in Euclidean space we obtain the point representation by calculating the envelope (cf. (2.8)).

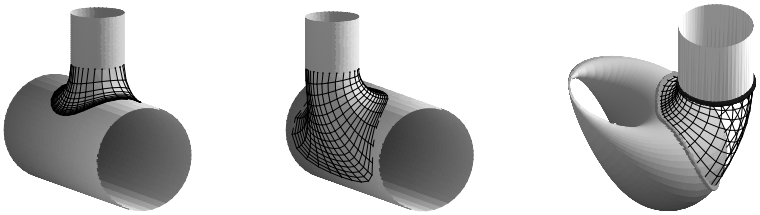


FIG. 8. *Various branching blends between cylinders and Dupin cyclides.*

If the dual data is of bidegree (d_t, d_u) then the bidegree of the solution (x_1, x_2, x_3) is $(3d_s - 2, 3d_t - 2)$ in general. Since $(d_s, d_t) = (2, 4)$ we can expect a solution of bidegree $(4, 10)$. Fortunately there exists a unique value of k in the expression of $\mathbf{n}_b$ (7.1) that enables us to drop bidegree down to $(3, 6)$.

It was proved in [21] that this is the minimal possible Laguerre invariant bidegree. The construction can be generalized to a few other positions of the given cylinders and then extended to any possible position by applying appropriate Laguerre transformations. Moreover by applying inversions similar PN branching blends can be generated between Dupin cyclides and cylinders or cones of bidegree $(4, 8)$. These possibilities are illustrated in Fig. 8.

8. Conclusions and open problems. We have given an overview of Pythagorean normal surfaces including their short history, an introduction to the language of Laguerre geometry, review of the most important classes of PN surfaces and their applications in geometric modeling. It has been shown that the dual approach in combination with universal rational parametrization ideas seems to be most promising not only for theoretical investigations but also for solving very practical modeling problems. There are still many questions that need to be discussed. Here are a few open problems for future research:

- Results on minimal parametrization degree of canal surfaces in Section 6.1 raises similar questions for other classes of PN surfaces. It is most important to understand possible reductions of degree when going from the dual representation to the point representation. Similar problems for PH curves were considered in [11].
- A free-form modeling scheme with PN surfaces without restrictions on the Gaussian curvature is a challenge. Perhaps the modeling scheme with parabolic cyclide patches (see Section 7.1) can be generalized using the universal rational parametrization of the Blaschke cylinder.
- Pieces of isotropic hypersurfaces of PN surfaces (e.g. between a surface and its offset) in Minkowski space projected down to Euclidean space $\mathbb{R}^3$ give nice examples of 3D rational parametriza-

tions of solids. Sufficient non-degeneracy conditions of such parametrizations will be useful in modeling.

We hope that this survey will help interested researches and people from industry to get an adequate impression about achieved results and the state of art in investigations of PN surfaces and their modeling applications.

REFERENCES

[1] Chen F. (2003). *Reparametrization of a rational ruled surface using the μ-basis*, Computer Aided Geometric Design, Vol. **20**, pp. 11–17.

[2] Choi H.I., Lee D.S., and Moon H.P. (2002), *Clifford algebra, spin representation, and rational parametrizations of curves and surfaces*, Advances in Computational Mathematics **17**, 2002, pp. 5–48.

[3] Cho H.C., Choi H.I., Kwon H.-S., Lee D.S., and Wee N.-S., *Clifford algebra, Lorentzian geometry, and rational parametrizations of canal surfaces*, Computer Aided Geometric Design **21**, 2004, pp. 327–339.

[4] Cox D., Krasauskas R., and Mustaţă M., *Universal rational parametrizations and toric varieties*, Topics in Algebraic Geometry and Geometric Modeling, Contemporary Mathematics **334**, 2003, pp. 241–265.

[5] Degen W., *Cyclides*, in: G. Farin, J. Hoschek, M.-S. Kim (eds.) Handbook of Computer Aided Geometric Design, Elsevier Science, 2002, pp. 575–602.

[6] Dietz R., Hoschek J., and Jüttler B., *An algebraic approach to curves and surfaces on the sphere and other quadrics*, Computer Aided Geometric Design **10**, 1993, pp. 211–229.

[7] Farouki R.T., *Pythagorean–hodograph curves in practical use*, in: Geometry Processing for Design and Manufacturing (Barnhill R.E., ed.), SIAM, Philadelphia, 1992, pp. 3–33.

[8] Farouki R.T. (2002). *Pythagorean Hodograph curves*, in: Handbook of Computer Aided Geometric Design (Farin G., Hoschek J., Kim M.-S. eds.), Elsevier, 2002.

[9] Farouki R.T., *Pythagorean-hodograph curves: algebra and geometry inseparable*, Geometry and Computing, Vol. **1**, Springer, 2008.

[10] Farouki R.T. and Neff C.A., *Algebraic properties of plane offset curves*, Computer Aided Geometric Design **7**, 1990, 101–127.

[11] Farouki R.T. and Pottmann H., *Polynomial and rational Pythagorean-hodograph curves reconciled*, in: G. Mullineux (ed.), The Mathematics of Surfaces VI, Oxford Univ. Press, 1996, pp. 355-378.

[12] Hoschek J. and Lasser D., *Fundamentals of Computer Aided Geometric Design*, A.K. Peters, Wellesley, MA, 1993.

[13] Jüttler B. (1998). *Triangular Bézier surface patches with a linear normal vector field*, in: The Mathematics of Surfaces VIII, Information Geometers, pp. 431–446.

[14] Jüttler B. and Sampoli M.L., *Hermite interpolation by piecewise polynomial surfaces with rational offsets*, Computer Aided Geometric Design **17**, 2000, pp. 361–385.

[15] Kazakevičiūtė M., *Blending of natural quadrics with rational canal surfaces*, PhD Thesis, Vilnius University, 2005.

[16] Kazakevičiūtė M., *Classification of pairs of natural quadrics from the point of view of Laguerre geometry*, Lithuanian Mathematical Journal **45**, 2005, pp. 63–84.

[17] Kosinka J. and Jüttler B., *MOS surfaces: Medial Surface Transforms with Rational Domain Boundaries*, in: R. Martin, M. Sabin, J. Winkler (eds.), The Mathematics of Surfaces XII, Lecture Notes in Computer Science **4647**, pp. 245–262, 2007.

[18] Krasauskas R., *Universal parametrizations of some rational surfaces*, in: A. Le Méhauté, C. Rabut and L.L. Schumaker (eds.): Curves and Surfaces with Applications in CAGD, Vanderbilt Univ. Press, Nashville, 1997, pp. 231–238.

[19] Krasauskas R., *Bézier patches on almost toric surfaces*, in: Elkadi M., Mourrain B. and Piene R. (eds.), Algebraic Geometry and Geometric Modeling, Mathematics and Visualization Series, Springer, 2006, pp. 135–150.

[20] Krasauskas R., *Minimal rational parametrizations of canal surfaces*, Computing **79**, 2007, pp. 281–290.

[21] Krasauskas R., *Branching blend of natural quadrics based on surfaces with rational offsets*, Computer Aided Geometric Design **25**, (2008), pp. 332–341.

[22] Krasauskas R. and Kazakevičiūtė M., *Universal rational parametrizations and spline curves on toric surfaces*, in: Computational Methods for Algebraic Spline Surfaces, ESF Exploratory Workshop, Springer, 2005, pp. 213–231.

[23] Krasauskas R. and Mäurer C., *Studying cyclides using Laguerre geometry*, Computer Aided Geometric Design **17**, 2000, pp. 101–126.

[24] Krasauskas R. and Zubė S., *Canal surfaces defined by quadratic families of spheres*, Proceedings of the COMPASS II Workshop, 2007, pp. 138–150.

[25] Landsmann G., Schicho J., and Winkler F., *The Parametrization of Canal Surfaces and the Decomposition of Polynomials into a Sum of Two Squares*, Journal of Symbolic Computation **32**, 2001, pp. 119–132.

[26] Leopoldseder S. and Pottmann H. (1998). *Approximation of developable surfaces with cone spline surfaces*, Computer-Aided Design **30**, 571–582.

[27] Lü W., *Rational canal surfaces*, Technical Report No. 13, Institut für Geometrie, TU Wien, 1995.

[28] Lü W. and Pottmann H., *Pipe surfaces with rational spine curve are rational*, Computer Aided Geometric Design **13**, 1996, pp. 327–339.

[29] Lü W. (1994). *Rationality of the offsets to algebraic curves and surfaces*, Applied Mathematics **9** (Ser. B), 265–278.

[30] Lü W. (1996). *Rational parametrization of quadrics and their offsets*, Computing **57**, 1996, 135–147.

[31] Peternell M., *Rational parametrizations for envelopes of quadric families*, Ph.D. Thesis, Institute of Geometry, University of Technology, Vienna, 1997.

[32] Peternell M., *Developable surface fitting to point clouds*, Computer Aided Geometric Design **21**, 2004, 785–803.

[33] Peternell M. and Pottmann H., *Designing rational surfaces with rational offsets*, Fontanella F., Jetter K., and Laurent P.J. (eds.), 1996, pp. 275–286.

[34] Peternell M. and Pottmann H., *Computing rational parametrizations of canal surfaces*, J. Symbolic Computation **23**, 1997, pp. 255–266.

[35] Peternell M. and Pottmann H. (1998). *A Laguerre geometric approach to rational offsets*, Computer Aided Geometric Design **15**, 223–249.

[36] Peternell M., Odehnal B., and Sampoli M.L.. *On quadratic two-parameter families of spheres and their envelope*, Computer Aided Geometric Design **25**, 342–355, 2008.

[37] Pottmann H. (1994). *Applications of the dual Bézier representation of rational curves and surfaces*, in: Laurent P.J., LeMéhauté A., and Schumaker L.L., eds., Curves and Surfaces in Geometric Design, A.K. Peters, Wellesley, MA, 377–384.

[38] Pottmann H. (1995). *Rational curves and surfaces with rational offsets*, Computer Aided Geometric Design **12**, 175–192.

[39] Pottmann H. (1995). *Curve design with rational Pythagorean–hodograph curves*, Advances in Comp. Math. **3**, 147–170.

[40] Pottmann H., *Studying NURBS curves and surfaces with classical geometry*, in: Mathematical Methods for Curves and Surfaces, eds: M. Dæhlen, T. Lyche, and L.L. Schumaker, Vanderbilt University Press, 1995, pp. 413–438.

[41] Pottmann H. and Farin G. (1995). *Developable rational Bézier and B-spline surfaces*, Computer Aided Geometric Design **12**, 513–531.

[42] Pottmann H., Lü W., and Ravani B. (1996). *Rational ruled surfaces and their offsets*, Graphical Models and Image Processing **58**, 544-552.

[43] Pottmann H. and Peternell M. (1998). *Applications of Laguerre Geometry in CAGD*, Computer Aided Geometric Design **15**, 165–186.

[44] Pottmann H. and Wallner J., *Computational Line Geometry*, Springer-Verlag, 2001.

[45] Powell M.J.D. and Sabin M., *Piecewise quadratic approximation on triangles*, ACM Transactions on Mathematical Software **3**, 1977, pp. 317–325.

[46] Pratt M.J., *Cyclides in computer aided geometric design*, Computer Aided Geometric Design **7**, pp. 221–242.

[47] Pratt M.J., *Cyclides in computer aided geometric design II*, Computer Aided Geometric Design **12**, 1995, pp. 131–152.

[48] Sampoli M.L., Peternell M., and Jüttler B., *Exact parametrization of convolution surfaces and rational surfaces with linear normals*, Computer Aided Geometric Design **23**, 2006, pp. 179–192.

[49] Sottile F., *Toric ideals, real toric varieties, and the moment map*, Topics in Algebraic Geometry and Geometric Modeling, Contemporary Mathematics **334**, 2003, pp. 225–240.

[50] Srivinas Y.L. and Dutta D., *An intuitive procedure for constructing geometrically complex objects using cyclides*, Computer-Aided Design **26**, 1994, pp. 327–335.

[51] Ueda K., *Blending between right circular cylinders with parabolic cyclides*, Proceedings of the Geometric Modeling and Processing 2000, April 10–12, 2000, pp. 390–397.

A LIST OF CHALLENGES FOR REAL ALGEBRAIC PLANE CURVE VISUALIZATION SOFTWARE

OLIVER LABS*

Abstract. Recently, the visualization of implicitly given algebraic curves and surfaces has become an area of active research. Most of the approaches either use raytracing, subdivision or sweeping techniques to produce a good approximate picture of the varieties, sometimes by using hardware equipment such as graphics processing units.

We provide a list of equations of plane curves which may serve as a list of benchmarks for visualization software. In most cases, we give whole series of examples which yield equations for infinitely many degrees. Even for low degrees, there is currently no software which visualizes all examples correctly in real–time, so we call them challenges.

For most of the equations in our list, we are able to prove that they are at least close to the most difficult possible ones. For convenience, our list is also available in the form of a text file. Moreoever, the paper includes a brief introduction to some of the terminology from singularity theory for researchers from the computer graphics community because singularities appear frequently when treating complicated cases.

Key words. Real algebraic geometry, computational geometry, geometric modeling, plane curves, singularities, visualization, algorithms, benchmarks, challenges.

AMS(MOS) subject classifications. Primary 14H45, 14B05, 14P05, 14Q05.

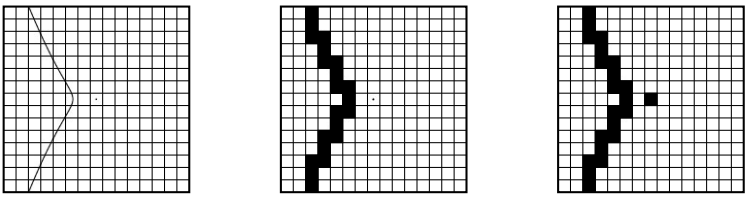
Introduction. A real algebraic plane curve of degree d in $\mathbb{R}^2$ is the zero–set of a possibly reducible polynomial of degree d in two variables with real coefficients; in this paper, we restrict ourselves to examples with rational, exact coefficients. We provide a list of equations of real plane curves which may serve as a list of benchmarks for visualization software. In most cases, we give whole series of examples which yield equations for infinitely many degrees. Dispite a lot of recent research activity, there is currently no software which visualizes all examples correctly in real–time, even for moderate degrees, so we call them challenges.

Most curves in our list are singular or are at least in some way related to singularities. After a short discussions of the term correct visualization, we thus give a brief introduction to singularity theory. Our examples which are presented in the remaining part of the article can roughly be divided into the following categories:

1. many or higher solitary points,
2. high tangencies at isolated singularities (e.g., A_k-singularities),
3. many or complicated isolated singularities,
4. many or complicated non–isolated singularities,
5. small deformations of one of the singular ones mentioned above.

*Universität des Saarlandes, Mathematik und Informatik, Gebäude E2.4, D-66123 Saarbrücken, Germany (Labs@Math.Uni-Sb.de, www.OliverLabs.net).

I.Z. Emiris et al. (eds.), *Nonlinear Computational Geometry*, The IMA Volumes in Mathematics and its Applications 151, DOI 10.1007/978-1-4419-0999-2_6,



(a) (b) (c)

FIG. 1. *(a) shows the plane curve* $e = y^2 + x^2(x+1)$ *(mind the solitary singular point in the center of the picture). (b) shows a bad visualization of it, (c) shows a correct visualization.*

The items above suggest that our examples will be difficult to visualize mainly because of their geometry and not just because of their coefficient size. This is indeed true; in particular, all our examples have rational coefficients and the number of digits needed to represent these is moderate. To keep the equations as simple as possible, many of our examples are actually reducible, i.e. they are products of several (often two) equations. Of course, in many cases, one could write down irreducible equations yielding similar visualization problems, but these would be a lot more complicated, so we do not include them here. To perform a fair comparison between several algorithms based on our examples, one should thus not allow a factorization prior to the main algorithm. Likewise, although the list does not try to produce dense polynomials in the sense that most monomials occur, this can often be realized by a simple coordinate change.

The list presented in this article will be updated on our website [Lab03] if problems become known which are more challenging than those given here. We will also provide the output of some of the existing visualization tools for each of the examples in our list in order to allow comparisons.

The first draft of this paper was written at the CMA at the University of Oslo. I thank all people there for their hospitality and the wonderful and inspiring atmosphere. In particular, I thank R. Piene for giving me the opportunity to stay at this great place for a month.

1. On correct visualizations of real algebraic plane curves. For simplicity, in this paper, we just ask for a correct image of the given plane curve up to pixel level: A pixel of the visualization area shall be colored in black if and only if the plane curve contains at least one point inside this pixel. This has the advantage that plane curves with solitary points are represented in a topologically correct way (see Figure 1). For our solutions given in this article, we always choose an area such that all singular points and all points with horizontal or vertical tangents can be seen in the picture.

The notion of correct visualization given above differs from another natural one which asks for a result reflecting the whole topology of the

plane curve correctly. E.g., if one of the unit squares of the visualization area contains two ***ovals***[1] inside each other then our notion of correct visualization above just shows one black pixel and does not mention the extra information about the geometry inside the pixel. In that case, the image shown thus won't be topologically correct.

1.1. An algorithm which produces correct visualizations. Before giving the challenges, we sketch a straightforward algorithm which is guaranteed to produce correct visualizations (up to pixel level) for a given equation of a plane curve C with rational coefficients as described above:

- Compute all zeros of C along the sides of the pixels (there are algorithms which assure correct output, e.g., using Sturm sequences), and draw a black pixel for all those squares containing a root.
- Compute all real points satisfying $C(x, y) = 0$ and $\frac{\partial C}{\partial y}(x, y) = 0$ up to a precision which ensures in which square the points lie (there are algorithms which yield certified results, e.g., Roullier's approach using Gröbner basis), and draw a black pixel for all those squares containing one of these points.

Of course, there are other algorithms which yield certified results and which may run faster. E.g., the website `http://exacus.mpi-inf.mpg.de` provides a web interface to the algorithm implemented in the Exacus library which even yields topologically correct output up to sub–pixel level.

1.2. The main visualization strategies. All known visualization methods with certified results are rather slow already from a moderate degree and a moderate number of pixels on. So, the challenge is to find an algorithm which works both correctly and fast.

Many algorithms try to use inexact computations for speed ups. For such algorithms, it is challenging to increase the number of curves which can be visualized correctly. Our list is an attempt to give developers of exact algorithms a tool to test the efficiency and to give developers of inexact algorithms a tool to test the correctness of their methods.

To understand the main difficulties, it is important to understand the basic visualization strategies which are currently used. A large class of algorithms first computes points (or at least an approximation of their x-coordinates) with vertical or horizontal tangents (which include the singular points) and then uses this information in order to ensure the topology of the output, e.g. using sweeping techniques. The initial step is often quite time–consuming; our list provides examples for which the number of singular points or points with vertical tangents is high.

Other algorithms proceed by subdividing the screen recursively either until the topology within a small rectangle can be guaranteed or until some

[1] An oval is a smooth connected component of a real plane curve which is intersected by any other plane curve in an even number of points (counting multiplicities).

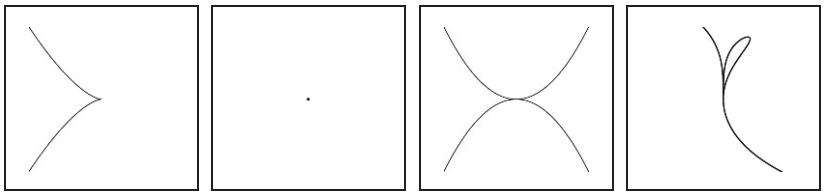


FIG. 2. *Plane curves with isolated singularities at the origin $O = (0, 0)$, (a) is an A_2-singularity or cusp, (b) is a higher solitary point, (c) is an A_3-singularity or tacnode, (d) is a more complicated singularity, a J_{13}-singularity to be precise.*

heuristic criterion tells the algorithm not to search deeper. Such algorithms usually have difficulties near singular or almost singular situations because there is still no efficient criterion known which says that a given rectangle does or does not contain exactly one singular point. Some heuristics work fine at very simple singularities, but most fail at higher singularities or almost singular points. Thus, our list also provides examples with complicated singularities and points which are almost singular in some sense.

2. Some notions from singularity theory. As mentioned earlier, many of our challenges will be related to singularities. In this section, we thus provide a short introduction to the necessary notions: A ***critical point*** of a real plane curve $f \subset \mathbb{R}^2$ of degree d is a point $p \in \mathbb{R}^2$ s.t. $\frac{\partial f}{\partial x}(p) = \frac{\partial f}{\partial y}(p) = 0$. If in addition $f(p) = 0$ then p is called a ***singular point*** or ***singularity*** (see also Figure 2). f is called ***smooth*** if it does not contain any singular point. p is an ***isolated singularity*** of f if there is an open neighborhood $V \subset \mathbb{R}^2$ of p which does not contain any other critical point. p is called ***non–isolated*** if this is not the case; for plane curves, this simply means that p is contained in a multiple factor of f, i.e. f is of the form $f = g^k \cdot h$ with $k \geq 2$ and $\deg(g) \geq 1$ with $p \in \{g = 0\}$. A ***solitary point*** is an isolated singularity p s.t. there is an open neighborhood $V \subset \mathbb{R}^2$ of p which does not contain any other point of the plane curve. Singularities have been classified in many respects. We only mention a few cases and refer to [AGZV85, BK86, Dim87] for details.

The ***multiplicity*** $\mathrm{mult}_p(f)$ of a singularity $p \in \mathbb{R}^n$ of a plane curve f is simply the degree of the lowest order term of f after translation of p to the origin. E.g., for the plane curves $x^2 - y^2$, $x^2 + y^2$ and $xy - y^5$, the origin $(0, 0) \in \mathbb{R}^2$ has multiplicity 2, i.e. it is a ***double point***; for $x^3 + y^3 + x^4 \cdot y^3$, the origin is a ***triple point***, etc. This definition makes sense because a generic line $l(t)$ through a k-tuple point p intersects the curve $f(x, y)$ with multiplicity k at p.

However, the multiplicity is only a very rough mean of classification of the singularities which can occur on plane curves; there are several nat-

ural ways of detailing further. E.g., after translation of a point p to the origin, the ***tangent cone*** $tangcone_p(f)$ of f at a point p is the sum of all terms of the lowest degree. An isolated singularity p of multiplicity m of a plane curve is called ***ordinary*** if and only if $tangcone_p(f)$ factors over the complex numbers into m different straight lines. E.g., for $x^2-y^2-x^3=0$, the origin is an ordinary double point as well as for $x^2+y^2-x^3=0$ (two complex conjugate lines intersect at the origin), and for $xy(x+y)-y^4=0$, the origin is an ordinary triple point.

But there are many double points apart from the ordinary ones. E.g., a singularity of a plane curve is called an A_k-singularity (or ***singularity of type*** A_k) if it can be written after a coordinate change which is locally a diffeomorphism in the form: $x^{k+1} \pm y^2 = 0$. Often, one specifies the sequence of signs together with the name of the singularity because many of the corresponding singularities differ essentially. E.g., $x^2+y^2=0$ and x^2-y^2 both define A_1-singularities in the complex classification, over the reals the first is of type A_1^+ (sometimes denoted by $A_1^\bullet$ because it is topologically a ***solitary point***), the second is of type A_1^-. The A_k-singularities are only the very beginning of a classification of singularities developed by Arnold, see section 8 and [AGZV85] for more information.

The index k appearing in the name A_k is the so-called ***Milnor number*** $\mu(f)$ of a singularity f of that type. In general, for a hypersurface singularity f at the origin, it is defined as the dimension of the ***Milnor algebra***:

$$\mu(f) := \dim_{\mathbb{Q}} \mathbb{Q}\{x_1,\dots,x_n\} \Big/ \left\langle \frac{\partial f}{\partial x_1},\dots,\frac{\partial f}{\partial x_n} \right\rangle,$$

where $\mathbb{Q}\{x_1,\dots,x_n\}$ denotes the ring of convergent power series and $\langle \dots \rangle$ denotes the ideal generated by the given polynomials. This dimension can be computed easily using local variants of Gröbner bases, so-called standard bases [GPS06]. It can also be computed in general purpose computer algebra systems by computing all points which vanish on all the partial differentials mentioned above: the Milnor number of f at the origin is then simply the multiplicity of the origin. For an A_k-singularity with equation $f = x^{k+1}+y^2$, we find $\partial f/\partial x = (k+1)x^k$, $\partial f/\partial y = 2y$. Thus a bases for the Milnor algebra is given by $1, x, \dots, x^{k-1}$ which shows that $\mu(f)=k$. There is a nice topological definition of the Milnor number, but we do not have enough place to describe that here. The ***Tjurina number*** $\tau(f) := \dim_{\mathbb{Q}} \mathbb{Q}\{x_1,\dots,x_n\}/\langle f, \frac{\partial f}{\partial x_1},\dots,\frac{\partial f}{\partial x_n}\rangle$ is another important invariant which measures the complexity of a singularity in a way similar to the Milnor number. In most of our examples, the Tjurina number actually equals the Milnor number; e.g., in the case of the A_k singularities mentioned above. However, for more complicated singularities this is usually not the case.

From a visualization point of view, one of the most basic invariants of a real isolated singular point of a plane curve is its number of ***halfbranches***:

In a neighborhood $V \subset \mathbb{R}^2$ of such a point p of a plane curve C, the set of points of C in $V \backslash \{p\}$ consists of a finite number of connected components, each of which is topologically a segment if V is small enough (see [BCR98, prop. 9.5.1]). The plane curves (a)–(d) shown in Figure 2 have 2, 0, 4, and 4 halfbranches at the singular point. Halfbranches always occur in pairs of two which together form the complex ***branches***. Of course, some singularities have more branches than twice the number of real halfbranches, e.g., $x^2 + y^2 = 0$ has two complex branches which are complex conjugate to each other, but no real halfbranch.

3. Solitary points. We already encountered the problem of visualizing solitary points (see, e.g., figure 1). Our first set of problems thus gives examples of plane curves with such features.

3.1. Many solitary points. Visualizing a plane curve with exactly one singularitiy at the origin correctly is significantly easier than the more general case with many singularities. The reason for this is that we can use computer algebra methods to analyze the local structure of the singularity at the origin. However, once we have many singularities, it is not easy to perform this analysis efficiently. This is why we give examples of plane curves with many singularities — and in this first case ordinary solitary double points ($A_1^\bullet$ singularities).

The maximum possible number of solitary points on a plane curve of degree d, denoted by $\mu^2_{A_1^\bullet}(d)$, is the genus $g(d)$ of a smooth plane curve if $d \neq 2, 4$:

$$\mu^2_{A_1^\bullet}(d) = g(d) = \frac{1}{2}(d-1)(d-2), \text{ if } d \neq 2, 4. \tag{3.1}$$

Because of their relation to ***smooth Harnack curves*** (i.e., curves which have the maximum possible number of connected components, see section 4) those curves which attain this bound are called ***rational Harnack curves***. Shustin already provided a construction of such curves for each degree d using his singular version of Viro's patchworking method [Shu98]. But his construction is not easy to perform explicitly, so for our list we take the rational Harnack curves which were constructed in [KO06] and call them $KO_d(x, y)$. The equations of the curves $KO_d(x, y)$ can be obtained as follows: Define the polynomials

$$f_d(x, y) := \prod_{k,l=0}^{d-1} \left(e^{\frac{2k\pi i}{d}} x + e^{\frac{2l\pi i}{d}} y + 1 \right).$$

One can show that there exists a real polynomial KO_d of degree d with:

$$KO_d(x^d, y^d) = f_d(x, y). \tag{3.2}$$

For $d \neq 2, 4$, these $KO_d(x, y)$ have exactly $\mu_{A_1^\bullet}(d) = g(d)$ solitary points which is the maximum possible number; in addition, they have exactly one smooth real branch not containing any singular point.

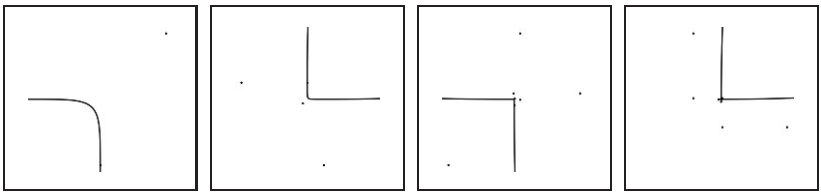


FIG. 3. *Some plane curves KO_d (see challenge 1). For $d = 6$, it is not possible to distinguish some of the $\frac{1}{2}(d-1)(d-2)$ solitary points in our visualization.*

CHALLENGE 1. *Visualize the plane curves*

$$KO_d(x, y) = 0$$

for $d = 4, 5, 6, \ldots$.
Solution: A correct visualization has to show those of the $\frac{1}{2}(d-1)(d-2)$ solitary points which lie in the rectangular area R in which we want to visualize the curve. As always, we choose R such that all singular points and all points with vertical or horizontal tangents are contained in it. Figure 3 shows four examples. □

3.2. Higher solitary points. The examples mentioned in the previous challenge are the most difficult possible ones in the sense that they reach the maximum possible number of solitary points. The solitary points in those examples are ordinary double points ($A_1^{\bullet}$-singularities), and can locally be written in the form $x^2 + y^2 = 0$. However, solitary points can become a lot more involved than just ordinary double points which happen to be solitary. E.g., the zeroset of the polynomials

$$SP_{k,l} := x^{2k} + y^{2l}, \quad k, l \in \mathbb{N}, \tag{3.3}$$

obviously only consists of the origin. To understand the difficulties which numerical visualization approaches might have with visualizing such solitary points, let us look at an example.

EXAMPLE 1. *For a real polynomial function of degree d in one variable, the greatest possible multiplicity of a root is d, realized by the example x^d. In more variables, an analogue to this is the following example: $f = x^d + y^d$. Here, $f|_{x=0} = y^d$ also has a root of multiplicity d (as well as $f|_{y=0} = x^d$).*

However, other plane curve singularities may have much higher orders of convergence towards zero along some specifically chosen curve. To show this, let us consider the plane curve:

$$g = (y - x^k)^2 + y^{2k}, \ k \in \mathbb{N}.$$

A diffeomorphic coordinate change φ which maps the origin to itself clearly does not change the multiplicity structure at the origin. We take

$$\varphi \colon \mathbb{R}^2 \to \mathbb{R}^2, \ (x, y) \mapsto (x, y - x^k)$$

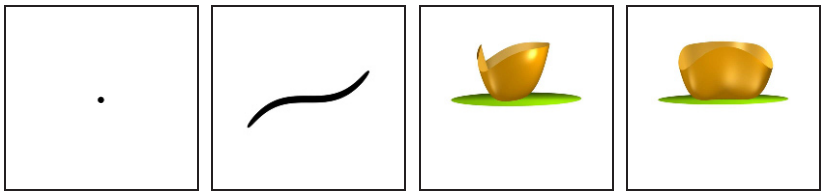


FIG. 4. *From left to right: A good and a bad visualization (considering a small function value as zero) of the plane curve* $g(x,y) = (y-x^k)^2 + y^{2k}, k = 2$, *and the graph of this function seen from two different directions.*

which is a diffeomorphism with jacobian matrix

$$J_\varphi = \begin{pmatrix} 1 & k \cdot x^{k-1} \\ 0 & 1 \end{pmatrix}.$$

This yields the new polynomial

$$(g \circ \varphi^{-1})(x,y) = y^2 + x^{2k^2} + r(x,y),$$

for some $r(x,y)$*, and one can show that there exists another diffeomorphism which transforms away the* $r(x,y)$*. In total, we find that* g *has the same multiplicity structure as* $y^2 + x^{2k^2}$*. The example of this plane curve* g *illustrates a well-known trick which produces plane curves with high Milnor number; variants also exist for other types of singularities, see e.g. [Wes05].*

Singularities which are related via such a diffeomorphic coordinate change are called **right–equivalent***. In these terms, at the origin the plane curve* g *is right–equivalent to* $y^2 + x^{2k^2}$*, the origin of* g *is thus called* A_{2k^2-1} *singularity. E.g., for* $k = 3$*,* g *has degree* 6 *and its singularity at the origin is right–equivalent to* $\widetilde{g} = y^2 + x^{2\cdot 3^2}$ *which defines an* A_{17}*-singularity. So, although the degree of* g *is only* 6*, it behaves at the origin basically as the polynomial* $\widetilde{g} = y^2 + x^{18}$ *of degree* 18*.*

To understand the numerical issues involved here, we look at the points $(\varepsilon, 0)$*, which have distance* ε *from the origin,* $\widetilde{g}$ *takes the values* $\widetilde{g}(\varepsilon,0) = \varepsilon^{18}$*. Figure 4 illustrates this; it shows two visualizations of our original plane curve* $g = (y-x^k)^2 + y^{2k}$ *for the case* $k = 3$*, (the left two images) together with its graph in* $\mathbb{R}^3$ *(the right two images). The leftmost image shows a point which is correct (the only solution to* $g = 0$ *is the origin). However, a bad numerical algorithm might produce a picture similar to the second one because* $g(x,y)$ *is very close to zero (in the second picture,* $|g(x,y)| < 10^{-3}$*) inside the black s-shaped area. The graph of the function* $g(x,y)$ *illustrates this: seen from one direction it looks basically like a parabola, but seen from some other direction we see that the function is a lot flatter — from our algebraic computations above, we know that it is basically of the form* $x \mapsto x^{17}$ *along the s-shaped area.*

To formalize the observations from the previous example, we introduce the following notion which measures the order at which the graph of a function $f(x, y)$ tends to zero around a solitary point:

DEFINITION 3.1. *Let $f(x, y)$ be a real plane curve with a solitary point at $p \in \mathbb{R}^2$. We denote by $C_p(\varepsilon)$ the circle around p with radius ε. Let*

$$m_{p,f}(\varepsilon) := \min\{|f(a, b)| \text{ s.t. } (a, b) \in C_p(\varepsilon)\}.$$

Then for $\varepsilon \to 0$, $m_{p,f}(\varepsilon) = O(\varepsilon^z)$ for some $z \in \mathbb{Q}$. We call $zerotang_f(p) := z$ the **zero–tangency** *of f at p.*

EXAMPLE 2. *We compute the zero–tangency for some simple plane curve singularities and also for the singularity already considered in the previous example. The latter one shows that the zero–tangency can be quadratic in the degree. This is the property which makes it particularly difficult for numerical algorithms to distinguish such solitary points from small connected components of plane curves:*

1. *$f = x^2 + y^2$, $p = (0, 0)$. Obviously, $zerotang_f(p) = 2$.*
2. *$f = x^2 + y^{2k}$, $k \geq 1$ $p = (0, 0)$. Then, $zerotang_f(p) = 2k$.*
3. *$f = x^{2k} + y^{2k}$, $k \geq 1$, $p = (0, 0)$. Then, $zerotang_f(p) = 2k$.*
4. *$f = (y - x^{2k})^2 + y^{4k}$, $k \geq 1$ $p = (0, 0)$. Then, $zerotang_f(p) = 2k^2$ because f is an $A^{\bullet}_{2k^2-1}$ singularity as we have seen in example 1.*

The most interesting curve in this example is the one which already occurred in example 1; we take a closer look at the singularities occurring there: It might be astonishing that for almost all degrees d, the maximum number $k(d)$ such that there exists a singularity of type $A^{\bullet}_{k(d)}$ on a plane curve of degree d is not known, even over the field of complex numbers. Currently, the best known upper and lower bounds are (see [GZN00], the asymptotic upper bound already follows from Varchenko's spectral bound [Var83], the upper bound in the cited paper is only slightly better):

$$\frac{15}{28}d^2 \lessapprox k(d) \lessapprox \frac{3}{4}d^2.$$

The authors of the cited paper give examples in degrees $d = 28 \cdot s + 9$ for $s = 0, 1, 2, \ldots$. This is not convenient for our list for which we need examples in low degree. So, we follow the idea presented in the example above:

CHALLENGE 2. *Visualize the plane curves*

$$SP_{k,l} := x^{2k} + y^{2l},$$

for $k = 1, 2, 3, \ldots$ and $l = 1, 2, 3, \ldots$, and

$$f^2_{k,l,+}(x, y) := (y - x^k)^l + y^{k \cdot l} \tag{3.4}$$

for $k = 2, 4, \ldots$ and $l = 2, 4, 6, \ldots$.

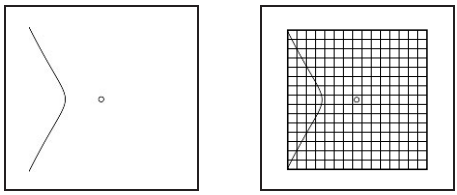


FIG. 5. *One way to try to visualize algebraic plane curves is to compute all zeros along some grid. This may miss some important features of the curve, as in the picture.*

Solution: These examples only contain a single solitary point. There is thus no need for a figure. However, their zero–tangencies differ quite a bit: $zerotang_{SP_{k,l}}((0,0)) = \deg(SP_{k,l})$, but $zerotang_{f^2_{k,l,+}}((0,0)) = l{\cdot}k^2$ which is quadratic in the degree for a fixed value of l. □

4. Smooth curves with many components. In many respects, the solitary points from the previous section are similar to small connected components of real plane curves. Harnack already showed in the 19$^{\text{th}}$ century [Har76] that a smooth plane curve cannot have more than $g(d)+1$ connected components where $g(d) = \frac{1}{2}(d-1)(d-2)$ is the genus of a (and thus any) smooth algebraic plane curve of degree d. Moreover, he was also able to prove that this number can actually be achieved which showed that the maximum possible number $b_0^2(d)$ of connected components on a smooth plane curve of degree d satisfies:

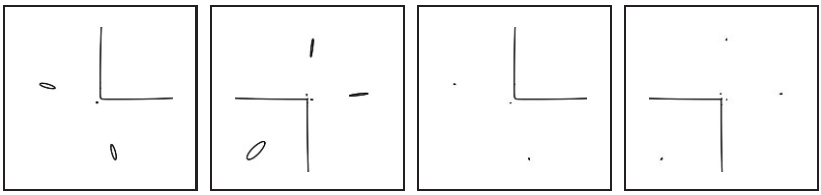
$$b_0^2(d) = g(d) + 1. \tag{4.1}$$

So, from this point of view the most difficult visualization challenge will be smooth curves of degree d having exactly $g(d)+1$ connected components.

4.1. Harnack curves. However, when considering the known algorithms for visualizing algebraic plane curves within a given rectangular area, we have to be a bit more specific. E.g., a simple algorithm chooses a certain number of horizontal and vertical lines (figure 5), computes the points of the curve on these lines, and draws these. If one of the connected components is entirely contained in one of the small rectangles formed by the sets of lines as in the figure, then this naïve method will not draw this component at all. As already mentioned, there are ways to solve this problem, but these are usually quite time–consuming.

The first trivial examples of plane curves which produce the problems indicated above are $x^2+y^2-\varepsilon = 0$ or $y^2+x^2(x+1)-\varepsilon = 0$ for some small ε. In higher degrees the most complicated examples are those which have the maximum possible number $b_0^2(d)$ of small ovals.

Such examples can be provided in the following explicit constructive way: Let $KO_d(x,y)$ be the rational Harnack curve of degree d given in (3.2) which have exactly $g(d)$ solitary points which is the maximum possible number. In addition, they have exactly one component which is not an



$\widetilde{KO}_{4,0.1}$ $\widetilde{KO}_{5,0.1}$ $\widetilde{KO}_{4,0.01}$ $\widetilde{KO}_{5,0.01}$

FIG. 6. *Some curves* $\widetilde{KO}_{d,\varepsilon}(x,y)$ *for* $d = 4, 5$ *for* $\varepsilon = 0.1,\ 0.01$ *(see challenge 3).*

oval, i.e. it is a pseudo–line. Brusotti's theorem (see, e.g., [BR90]) tells us that we can deform each of these $g(d)$ solitary points into small ovals. An explicit way of doing this is the following (see [Cos92] for a more general case):

$$\widetilde{KO}_{d,\varepsilon}(x,y) := KO_d(x,y) + \varepsilon x \cdot \frac{\partial KO_d(x,y)}{\partial x} + \varepsilon y \cdot \frac{\partial KO_d(x,y)}{\partial y} \qquad (4.2)$$

for some small $\varepsilon > 0$.

CHALLENGE 3. *Visualize the plane curves*

$$\widetilde{KO}_{d,\varepsilon}(x,y)$$

for $\varepsilon = 10^{-i}$ *for* $i = 1, 2, 3, \ldots$ *and* $d = 2, 3, \ldots$.
Solution: The visualization has to show all the $g(d) = \frac{1}{2}(d-1)(d-2)$ ovals if the rectangular area in which we want to see the curve is large enough. If ε is too big then some of the ovals might join. We give a few explicit cases in Figure 6. □

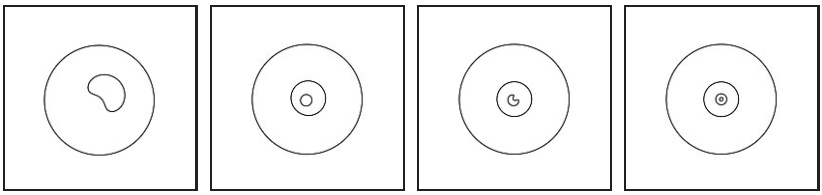
4.2. Nested ovals. Some of the ovals might be ***nested***, i.e. one oval is contained in one or more others. It is still an open question which topological configurations of ovals are actually possible for a real algebraic plane curve. Hilbert already knew that there have to be some restrictions, and he posed the question on these configurations as his 16^{th} problem.

As mentioned in the introduction, most of the time, we restrict ourselves to the easier task of just finding one point within each unit square in the rectangular area of interest. But as some algorithms try to produce topologically exact results, we give some simple examples of curves with nested ovals:

CHALLENGE 4. *Visualize the plane curves*

$$Nest^2_{d,\varepsilon,k} := (x+y) \cdot \varepsilon^{k+\lfloor d/2 \rfloor} + \prod_{j=1}^{\lfloor d/2 \rfloor} (x^2 + y^2 - \varepsilon^j) \qquad (4.3)$$

for $d = 2, 3, \ldots$, $k = 0, 1, 2, 3, \ldots$ *and some small* $\varepsilon > 0$, *e.g.* $\varepsilon = 10^{-i}$ *for* $i = 1, 2, 3, \ldots$.



$Nest^2_{6,0.1,1}$ $Nest^2_{6,0.1,2}$ $Nest^2_{8,0.1,4}$ $Nest^2_{8,0.1,6}$

FIG. 7. *Some curves* $Nest^2_{d,\varepsilon,k}$ *(see challenge 4).*

Solution: For $\varepsilon > 0$ small enough or $k \geq 0$ big enough, the curve consists of exactly $\lfloor d/2 \rfloor$ nested ovals which is the maximum possible number according to Bézout's theorem. In Figure 7, we show both cases. □

4.3. Small non–real ovals. To numerical visualization tools, a curve which is locally of the form

$$SP_{k,l,\varepsilon} := x^{2k} + y^{2l} + \varepsilon, \quad k, l \in \mathbb{N}, \varepsilon > 0, \tag{4.4}$$

poses serious problems if $\varepsilon > 0$ is small enough. A software which performs exact computations will detect that these polynomials do not have any real root. But a numerical software might draw one or more pixels because the function $(x, y) \mapsto x^{2k} + y^{2l}$ is very close to zero even far from the origin, in particular for large values of k, l. Similarly, we can take the plane curves $f^2_{k,l,+}(x, y) = (y - x^k)^l + y^{k \cdot l}$ of challenge 2 having higher solitary points and deform them slightly.

CHALLENGE 5. *Visualize the plane curves*

$$SP_{k,l,\varepsilon} = x^{2k} + y^{2l} + \varepsilon$$

for $k, l \in \mathbb{N}$*, and*

$$f^2_{k,l,+,\varepsilon}(x, y) := (y - x^k)^l + y^{k \cdot l} + \varepsilon \tag{4.5}$$

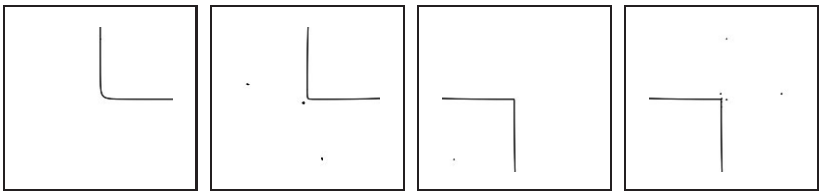
for $k = 1, 2, 3$*,* $l = 2, 4, 6, \ldots$ *for some small* $\varepsilon > 0$*, e.g.* $\varepsilon = 10^{-i}$ *for* $i = 1, 2, 3, \ldots$.

Solution: All plane curves have no real point, so we do not show a figure. □

If we want to produce curves which have such a local feature at many points, we can use the curves $KO_d(x, y)$ with many solitary points and take small deformations of them. As some of the solitary points of $KO_d(x, y)$ are local maxima of the graph $\{z - KO_d(x, y) = 0\} \subset \mathbb{R}^3$ and some others are local minima, we always get some small ovals if we add or subtract a small constant ε from these equations, and also some non-real "ovals".

CHALLENGE 6. *Visualize the plane curves*

$$KO^-_{d,\varepsilon}(x, y) := KO_d(x, y) - \varepsilon \tag{4.6}$$



$KO^+_{4,0.1}$ $KO^-_{4,0.1}$ $KO^+_{5,0.1}$ $KO^-_{5,0.1}$

FIG. 8. *Some curves which are small deformations of the rational Harnack curves with the maximum possible number of solitary points. Some of the solitary points are deformed into small ovals (appearing as dots in the pictures), some have disappeared into the complex domain (see challenge 6).*

and

$$KO^+_{d,\varepsilon}(x,y) := KO_d(x,y) + \varepsilon \tag{4.7}$$

for $d = 3, 4, \ldots$ *and for some small* $\varepsilon > 0$, *e.g.* $\varepsilon = 10^{-i}$ *for* $i = 1, 2, 3, \ldots$.
Solution: The curves KO_d have one connected component which is not an oval, i.e. a pseudo–line. In a neighborhood of this component, the function $KO_d(x,y)$ takes positive values on one side and negative values on the other. This causes all those solitary points of KO_d to disappear which are on one of the sides of the pseudo–line, depending on the sign in front of the ε of $KO^\pm_{d,\varepsilon}(x,y)$. The others become small ovals if ε is small enough.□

5. High tangencies at isolated singularities. In this section, we restrict ourselves to singularities with up to two complex branches and thus up to four real halfbranches. Later, we will mention some curves which have more complicated singularities. To produce equations yielding curves of low degrees with few halfbranches and high tangencies, we use again the trick mentioned in section 3.2 which yields A^-_k-singularities for a high Milnor number k. As mentioned before, the maximum possible number $k(d)$ such that there exists a singularity of type $A_{k(d)}$ on a plane curve of degree d is not known in most cases and we use the plane curves $f^2_{k,l}(x,y)$ which yield high Milnor numbers.

CHALLENGE 7. *Visualize the plane curves*

$$f^2_{k,l,-}(x,y) = (y - x^k)^l - y^{k\cdot l} \tag{5.1}$$

for $k = 2, 3, \ldots$ *and* $l = 2$.
Solution: The examples have two pairs of real halfbranches which are hard to distinguish close to the singular point, see Figure 9. For subdivision visualization algorithms which are based on rectangular grids, these curves are very hard to visualize correctly in an efficient way. □

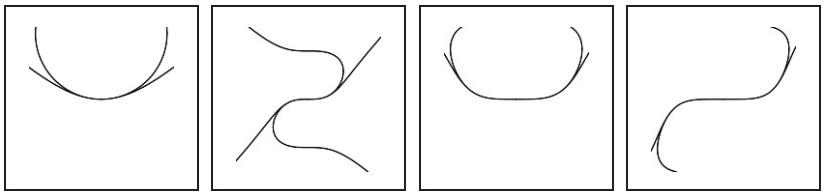


FIG. 9. *The plane curves* $f^2_{k,l,-}(x,y)$ *for* $l=2$ *and* $k=2,3,4,5$ *which have degree* $4,6,8,10$ *and singularities of type* A_7, A_{17}, A_{31}, A_{49} *(see challenge 7).*

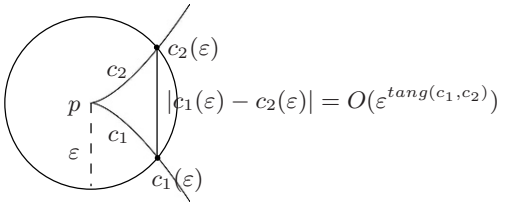


FIG. 10. *The definition of tangency of two halfbranches* c_1, c_2 *of a plane curve at an isolated singularity* p.

To make the problem coming from the halfbranches which are close to each other more precise, we introduce the notion of ***tangency*** of an isolated singularity p of a real plane curve f as follows[2]:

DEFINITION 5.1. *Let* $c_i, i=1,2,\ldots,N$ *be the real halfbranches of a real plane curve* f *at a singular point* p. *Then the* **tangency** $tang_f(p)$ *of* f *at* p *is the maximum of the tangencies of the* $\binom{N}{2}$ *pairs of halfbranches* c_i *where the tangency* $tang(c_i,c_j)$ *between two such components* $c_i, c_j, i\neq j$, *is defined as follows.*

If the c_i *and* c_j *have different tangent directions at* p *then* $tang(c_i,c_j)=1$. *If* c_i *and* c_j *have the same tangent direction* t *at* p, *but* c_i *is on one side of the perpendicular to* t *through* p *and* c_j *on the other then* $tang(c_i,c_j)=0$.

In the remaining case (see Figure 10) let $c_i(\varepsilon) := c_i \cap C_{(0,0)}(\varepsilon)$ *be the point of* c_i *at distance* ε *from* p *(we choose* $\varepsilon>0$ *small enough, s.t. this intersection is unique). Then for* $\varepsilon\to 0$, $|c_i(\varepsilon)-c_j(\varepsilon)|=O(\varepsilon^{tang(c_i,c_j)})$ *for some rational number* $tang(c_i,c_j)\in\mathbb{Q}$. *We call* $tang(c_i,c_j)$ *the* **tangency of** c_i **and** c_j. *The* **tangency of the singularity** p *is the maximum tangency of all pairs of real halfbranches. We denote the* **tangency of a real plane curve** f *at* p *by* $tang_p(f)$.

[2]Similar notions have already appeared in the literature, e.g. in Arnold's overview paper from 1968 on singularities of smooth mappings, c.f. [Arn81, p. 3–45].

As a diffeomorphism is locally a linear isomorphism, two singularities which are right–equivalent have the same tangency. The tangencies of normal forms of singular points are particularly easy to compute:

EXAMPLE 3.

1. *The tangencies of the two halfbranches c_1, c_2 of the normal forms $x^{k+1} + y^2$ of an A_k^--singularity, k even, are $tang(c_1, c_2) = \frac{k+1}{2} (= \frac{d}{2})$. Thus, the tangency of an A_k-singularity, k even, is:*

$$tang(A_k^-) = \frac{k+1}{2}.$$

2. *The tangencies of the plane curves $f^2_{k,2,-}(x, y)$ of degree $d = 2k$ from the challenge above are (as A_{2k^2-1}-singularities):*

$$tang_{(0,0)}(f^2_{k,2,-}) = \frac{2k^2}{2} = \frac{2d^2}{8}.$$

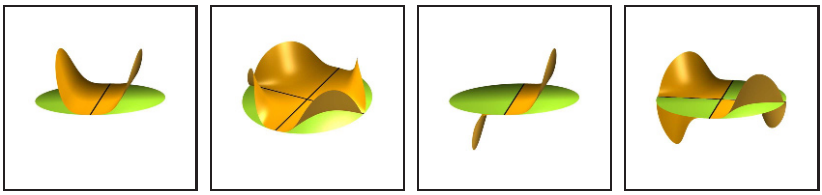
3. *The normal forms $x^{k+1} + y^2$ of an $A_k^\bullet$-singularity, k odd, are plane curves which only consist of a solitary double point. They have no real halfbranch, so their tangency is zero:*

$$tang(A_k^\bullet) = 0.$$

4. *The tangencies of the normal forms of the so-called singularities of type $E^\pm_{6k+1}$ with equation $y \cdot (y^{k-2} \pm x^2)$ are: $k - 2$ $(= d - 1)$ (for E^-_{6k+1}) and 1 (for E^+_{6k+1}).*

6. High tangencies at non–isolated singularities. In the case of plane curves, a non–isolated singularity is basically just a point which is contained in some multiple component of the curve. So, for plane curves non–isolated singularities are in principal not a big visualization issue because it is not difficult to compute the squarefree part of a polynomial, at least if the input data is exact.

However, if we allow small deformations of plane curves with multiple components (e.g., because we work with limited precision) then these components will no longer be multiple, but only approximately multiple. Good visualizations of such plane curves are not easy; in fact, it is not even clear what a good visualization of such a curve should be. Should it be an exact visualization of the deformed curve or should it be a visualization of the original one? If one is interested in the latter one, then one has to answer the quite non-obvious question what the original curve was when starting from a deformed one. In the case of non–isolated singularities, this is related to the question of computing approximate greatest common divisors which is a hard problem. Approximate isolated singularities are probably an even harder object of study.



x^4 $x^4 \cdot y^2$ x^5 $x^5 \cdot y^3$

FIG. 11. *Some graphs of plane curves with non–isolated singularities.*

The same difficulty comes up when working with only a small number of digits which has become quite common recently because of the invention of graphics cards with many processors. Rounding the coefficients of a plane curve with multiple components yields a curve which only has approximately multiple components.

These are some reasons to include examples of plane curves with non-isolated singularities. Such singularities can basically be divided into two classes: points on the one–dimensional component which have the same type of singularity as all points on the component in some small neighborhood, and the other points. The former are simply multiple components, the latter are multiple components which contain in addition a higher singular point on them.

A natural way to measure the type of singularity of the first kind of points (and thus all but finitely many points) is to take the type of the isolated singularity of a generic line section through the non–isolated singularity; this type is called ***transversal type*** of the non–isolated singularity. This yields a polynomial in a single variable; the highest possible multiplicity is achieved by x^d where d is the degree of the plane curve. Thus, the non–isolated singularity with the most complicated transversal type of singularity on a plane curve of degree d is given by $f(x, y) = x^d$.

To obtain examples of the more complicated type of non–isolated singularities, we combine the multiple line x^d with other multiple components or isolated singularities, some of which have the same tangent direction as the multiple component. To get an idea about the difficulty of the challenge, take a look at the graphs shown in Figure 11.

CHALLENGE 8. *Visualize the plane curves*

$$\begin{aligned} ni_m &:= x^m, \\ ni_{m,k} &:= x^m \cdot y^k, \\ nix_{m,n,k,l} &:= x^m \cdot (f_{k,l,-}^2(x, y))^n, \\ niy_{m,n,k,l} &:= y^m \cdot (f_{k,l,-}^2(x, y))^n, \end{aligned}$$

for $k, l, m, n = 2, 3, 4, \ldots$.

Solution: A correct visualization of the plane curves ni_m consists of a single vertical line; $ni_{m,k}$ are two lines which intersect in a right angle

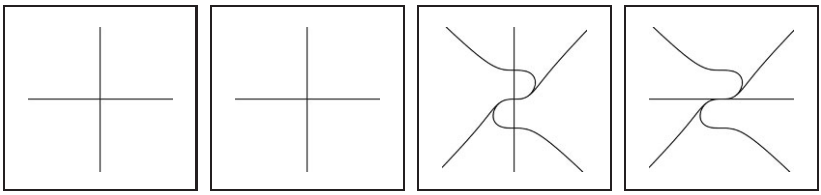


FIG. 12. *Visualizations of some non–isolated singularities: the components of the plane curve shown in the images have higher multiplicities (see challenge 8).*

(see Figure 12). $nix_{m,n,k,l}$ and $niy_{m,n,k,l}$ are the unions of a vertical, resp. horizontal, straight line, and the plane curve $f^2_{k,l,-}(x,y)$ (see challenge 7). This causes trouble to software which does not compute the squarefree part first or which only deals with few digits such that the multiple component does only appear as an approximately multiple component. □

7. Many isolated singularities. Plane curves with many singularities are plane curves with small genus, e.g. rational curves which are curves with genus zero. Similar to section 3.1 the difficulty in visualizing a plane curve with many singularities correctly is caused by the fact that it is not trivial to work with the coordinates of the singular points, in particular if they live in some higher extension of $\mathbb{Q}$.

In this section, we explain briefly a general and classically known method for globalizing the local structure of a singularity. Certainly, Shustin's singular version of Viro's patchworking method provides a very powerful method (at least in the case of plane curves) for doing this; however, it is not so easy to write down a list of explicit equations via this method. So, we use the classical strategy mentioned previously. Let us start by considering the A_2-singularity given by $f = x^3 - y^2$. Essentially, near the origin O, the curve f is the difference of the smooth curve $g\colon x = 0$ (tripled!) and another smooth curve $h\colon y = 0$ (doubled!) with the property that g and h have different tangent directions near the origin (i.e. g and h are ***transversal*** near O). As the type of a singularity only depends on the local behaviour of the curve, we can replace g and h by any other two plane curves which are smooth at O and which meet transversally at O.

In this way, we can easily produce plane curves with many real singularities which are locally of the type $x^k - y^l$:

CHALLENGE 9. *Visualize the plane curves*

$$C_{k,l} := (dfold_{\lfloor l/k \rfloor})^k - (x^2 + y^2 - 1)^l$$

for $2 = k \leq l$. *Here,*

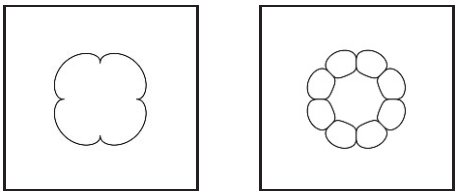


FIG. 13. *Some plane curves with many real singularities which all have higher tangencies (see challenge 9).*

$$dfold_d(x,y) := \prod_{j=1}^{d} \left(x \cdot \sin \frac{2\pi j}{d} + y \cdot \cos \frac{2\pi j}{d} \right)$$

are the d straight lines through the origin discussed in section 8.1.
Solution: The $\lfloor l/k \rfloor \cdot 2$ singularities $C_{k,l}$ at the intersections of the circle x^2+y^2-1 with the d straight lines are of the type $x^k - y^l$. For fixed k, the number of these singularities grows linearly in the degree $d = 2l$, and also the Milnor number of the singularity grows linearly in d. It is not difficult to show that this is asymptotically optimal, see Figure 13. □

There are many variants of the construction above. E.g., we can use the plane curves $f^2_{k,l,-}$ mentioned in challenge 7:

CHALLENGE 10. *Visualize the plane curves*

$$F^2_{k,l,m,-} := (y - ((x-1)\cdot(x-2)\cdots(x-m))^{\lfloor k/m \rfloor})^l - y^{k\cdot l}$$

for $l = 2$, $k = 2, 3, \ldots$, $2 \le m \le k$.
Solution: From the discussion above and challenge 7, it is clear that the plane curves $F^2_{k,2,m,-}$ have m singularities of type A_j for $j = \lfloor k/m \rfloor \cdot k \cdot l - 1 = 2k\lfloor k/m \rfloor - 1$ (see Figure 14). For a fixed number m of singularities, the Milnor number j grows quadratically in the degree $d = \frac{k}{2}$. Conversely, for fixed j, the number m of singularities grows quadratically in the degree. It is not difficult to show that this is asymptotically optimal. □

The coordinates of the singularities of the plane curves from the previous challenge are rational. As already mentioned, for some visualization algorithms, it is much harder to work with examples whose singularities have non-rational coordinates. To get such plane curves, we adapt the previous examples by replacing the product $(x-1)\cdots(x-m)$ by a function $f(x)$ which has only real roots none of which is rational for $\deg f$ even and one of which is rational for $\deg f$ odd. To give a concrete example, consider the so–called Tchebychev polynomial $T_m(x) \in \mathbb{R}[x]$ of degree m with critical values -1 and $+1$. This can either be defined recursively by $T_0(x) := 1$, $T_1(x) := x$, $T_m(x) := 2{\cdot}x{\cdot}T_{m-1}(x) - T_{m-2}(x)$ for $m \ge 2$,

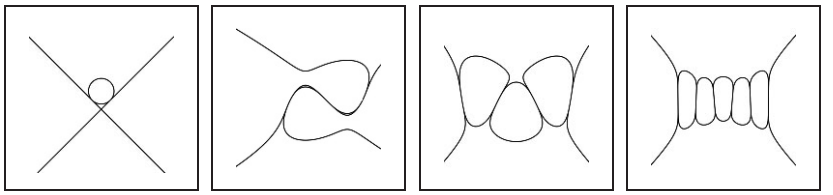


FIG. 14. *Some plane curves with many real singularities (see challenge 10):* $F^2_{2,2,2,-}$, $F^2_{4,2,3,-}$, $F^2_{4,2,4,-}$, $F^2_{6,2,6,-}$. *The figure also shows their degree and number and type of singularities on them.*

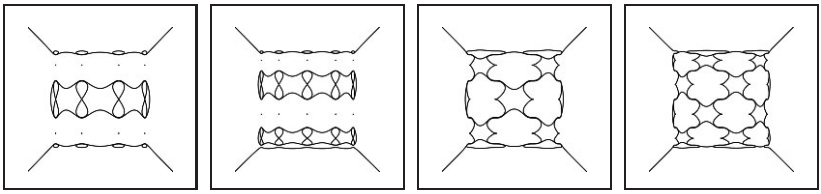


FIG. 15. *Some plane curves with many real higher singularities most of which have non–rational coordinates (see challenge 12).*

or implicitly by $T_m(\cos(x)) = \cos(mx)$. Most polynomials $T_m(x)$ have no rational root apart from 0.

CHALLENGE 11. *Visualize the plane curves*

$$FT^2_{k,l,m,-} := (y - T_m(x))^{\lfloor k/m \rfloor})^l - y^{k \cdot l}$$

for $l = 2$, $k = 2, 3, \ldots$, $2 \le m \le k$.
Solution: Essentially, these curves look similar to those in the previous challenge, so we do not show any pictures here. □

When also replacing the variable y in the previous challenge by the Tchebychev polynomial $T_m(y)$, we obtain curves whose number of higher singularities is at least m^2 while their coordinates are non–rational in most cases. In addition to the constructed higher singularities, some other singularities such as real nodes or solitary points may appear:

CHALLENGE 12. *Visualize the plane curves*

$$FTT^2_{k,m,-} := (T_m(y) - T_m(x)^k)^2 - T_m(y)^{2k},$$

for $k = 2, 3, \ldots$, $m = 2, 3, \ldots$.
Solution: We only give four examples in Figure 15: $k = 2, 3$, $m = 4, 5$. □

To obtain similar equations with only solitary points as real singularities at non–rational coordinates, one may use

$$FTT^2_{k,m,+} := (T_m(y) - T_m(x)^k)^2 + T_m(y)^{2k}. \tag{7.1}$$

We do not include a visualization of these curves because they only consist of higher solitary points whose coordinates are the roots of the T_m.

8. Complicated isolated singularities. What is a complicated singularity? This is not a common notion in singularity theory. Most singularities which occurred in this article up to now did not have more than four halfbranches at the singular points. For us, a complicated singularity should thus have more than 4 halfbranches. This criterion alone also applies to ordinary singularities (see the following section) which do not look very complicated. We thus also give examples of isolated singularities which are not ordinary and which have more than four halfbranches. Algebraically, it turns out that there are two other natural measures for the complexity of a singularity, namely the corank and the modality. We will dicsuss these notions further down.

8.1. Ordinary singularities. We already mentioned ordinary plane curve singularites in the introduction: An ordinary m-fold point is locally the intersection of exactly m different straight lines, in particular all complex branches have pairwise different tangent directions at the singular point. Ordinary singularities are very special. From the point of view of the tangency these are the easiest singularities because their tangency is by definition just 1.

Nevertheless, ordinary singularities are quite interesting. E.g., there is a number which is strongly related to these points, the so–called ***delta invariant***. We do not give a formal definition of this notion, but intuitively, a singularity with delta invariant δ concentrates δ ordinary double points. It is related to the Milnor number and the number r of complex branches through the singular point via the formula:

$$\mu = 2\delta - r + 1. \tag{8.1}$$

E.g., let us consider an ordinary 3-fold point (see Figure 16(a)). When moving one of the three lines slightly (figure 16(b)), we see that we find three ordinary double points in a small neighborhood of the original singularity. Similarly, if we take an ordinary m-fold point, we get $\binom{m}{2}$ ordinary double points when moving $m-2$ of the lines slightly in a generic way. A deformation of a singular point p which produces several singularities whose delta invariants sum up to the delta invariant of p is called a ***delta constant (or δ-constant) deformation***. According to equation (8.1), an ordinary m-fold point thus has Milnor number $2\cdot\binom{m}{2} - m + 1 = (m-1)^2$: The Milnor numbers of ordinary double, triple, quadruple points are $1, 4, 9$. An A_{2k}-singularity of a plane curve has only one branch and Milnor number $2k$; a δ-constant deformation into a curve with only ordinary double

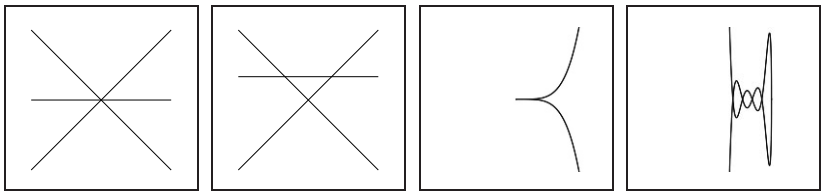


FIG. 16. *(a) an ordinary triple point, (b) a δ-constant deformation of an ordinary triple point, (c) an A_8-singularity, (d) a δ-constant deformation of an A_8-singularity (with δ-invariant 4).*

points thus has $2k/2 = k$ such singularities (see Figure 16). The δ invariant is a very important number because the genus of an irreducible plane curve f of degree d is

$$g(f) = \frac{(d-1)(d-2)}{2} - \sum_{s \text{ singularity of } f} \delta(s), \tag{8.2}$$

where $\delta(s)$ denotes the δ-invariant of the singularity s. Together with the fact $g(f) \geq 0$, this formula yields in particular an upper bound on the maximum possible number $\mu(s,d)$ of singularities of any fixed type s on a plane curve of degree d, namely: $\mu(s,d) \leq \frac{(d-1)(d-2)}{2\delta(s)}$. This number $\mu(s,d)$ is only known in very few cases.

CHALLENGE 13. *Visualize the plane curves*

$$dfold_d(x,y) = \prod_{k=1}^{d} \left(x \cdot \sin \frac{2\pi k}{d} + y \cdot \cos \frac{2\pi k}{d} \right)$$

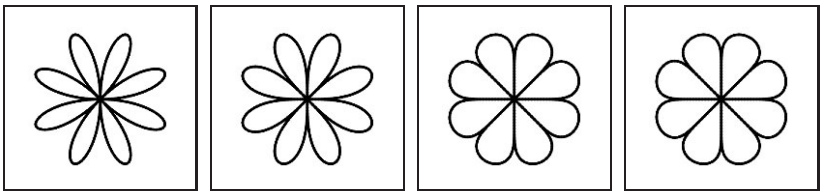
and

$$dfold^{fl}{}_d(x,y) = dfold_d(x,y) - (x^2+y^2)^{\lfloor d/2 \rfloor + 1}$$

for $d = 2, 3, 4, 5, 6, \ldots$.

Solution: The curves $dfold_d$ are just d-gon symmetric sets of d straight lines through the origin. There is thus no need for a figure. Notice that these polynomials actually do have rational coefficients; but it is more convenient to write them down with sin and cos. These curves are obviously those with the maximum possible number of halfbranches at a singularity of a plane curve of degree d.

The curves $dfold^{fl}{}_d$ have the same singularities, but are not factorizable into d straight lines and look like flowers similar to the curves shown in Figure 17 (but all branches have different tangent directions). □



$dfold_{4,1}$ $dfold_{4,2}$ $dfold_{4,3}$ $dfold_{4,4}$

FIG. 17. *The plane curves* $dfold_{k,l}$ *for* $k = 4$, $l = 1, 2, 3, 4$ *(see challenge 14) of degree* $10, 12, 22, 24$, *respectively.*

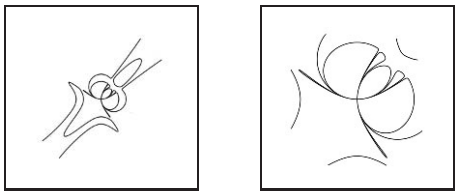


FIG. 18. *The plane curve* $fl4_2$ *of degree* 18 *at two different zoom levels.*

8.2. Some isolated singularities With many halfbranches. The curves $dfold_d$ with one ordinary d-fold singularity are those with the points with the highest number of halfbranches possible for a fixed degree. However, their tangency is quite low, namely 1. We now provide examples with arbitrarily high tangency and also an arbitrary number of halfbranches:

CHALLENGE 14. *Visualize the plane curves*

$$dfold_{k,l}(x, y) := (dfold_k(x, y))^2 - (x^2 + y^2)^{k+l}.$$

for $k = 2, 3, 4, 5, 6, \ldots$ *and* $l = 1, 2, 3, \ldots$.
Solution: These curves look like flowers, see Figure 17. □

It is not difficult to adapt the construction above to obtain similar plane curves with more than two halfbranches having the same tangent direction at the singular point. When using the trick from challenge 7 this produces plane curves with nice singularities. To give an example, we consider the plane curve

$$\begin{aligned} fl4_k := &(((y - x^k)^2 - y^{2k}) \cdot ((y - x^k)^2 - 2y^{2k})) \\ &\cdot(((x - y^k)^2 - x^{2k}) \cdot ((x - y^k)^2 - 2x^{2k})) \; - \; (xy)^{4k+1}, \quad k \geq 2. \end{aligned}$$

$fl4_k$ has exactly two tangent directions at the singular point and 8 halfbranches for each of them (figure 18).

8.3. Complicated singularities. As mentioned in the introduction to section 8, the term complicated singularity is not a commonly used one. We basically mean singularities which have not too small invariants such

as the number of branches, the corank, and the modality (see below for the definitions of these notions). The complicated singularities presented in this section are neither A_k-singularities nor ordinary singularities.

The ***corank*** of an algebraic plane curve f at a point p is defined as the corank of the Hessian matrix, i.e. the matrix of the second partial derivatives of f at p. The corank of a plane curve singularity is thus at most 2. Basically, the corank is the number of variables needed to define the type of the singularity up to right equivalence. More precisely, the generalized Morse Lemma tells us that for any hypersurface given by a polynomial $h(x_1, \ldots, x_n)$ with an isolated singularity at p there is a diffeomorphism which brings h into the form:

$$g(x_1, \ldots, x_c) \pm x_{c+1}^2 \pm \cdots \pm x_n^2,$$

where $c = \operatorname{corank}(h)$. From this it is clear that the corank is some rough measure for the complexity of the singularity. E.g., the ordinary double points $x^2 \pm y^2$ are the only plane curve singularities with corank 0, i.e. with a Hessian matrix of full rank.

Another measure for the complexity of a singularity is the so-called modality. We will not give a precise definition of this here, but we will just illustrate it using an example: Three lines in the plane through the origin can always be transformed into a specific set of three lines, e.g. x, y, $x - y$, by a local diffeomorphism. However, four lines cannot; indeed, the cross–ratio of the four slopes is an invariant. Thus, the class of ordinary singularities of multiplicity four (i.e. locally the intersection of four lines) depends on one parameter. And this is the number of moduli, called the ***modality*** which is 1 in the example. Singularities with modality zero are called ***simple singularities***. These are the singularities of types $A_k^{\pm}$ ($x^{k+1} \pm y^2 = 0$, $k \geq 1$), $D_k^{\pm}$ ($x^2 y \pm y^{k-1} = 0$, $k \geq 4$), $E_6^{\pm}$ ($x^3 \pm y^4 = 0$), $E_7^{\pm}$ ($x^3 \pm xy^3 = 0$), $E_8^{\pm}$ ($x^3 \pm y^5 = 0$). These singularities have many beautiful relations with other areas of mathematics. Here, we only mention Lie groups who gave their name to the corresponding singularities (see, e.g., [Dur79]) and regular polyhedra.

For our list of challenges we only give those singularities which have one modulus, called ***unimodal singularities***, and which are part of a series of equations for infinitely many degrees. For explicit equations of even more complicated singularities (e.g., with modulus 2 or corank 3), see [AGZV85, volume I, chapter 17]. The cited chapter also contains a list of some exceptional singularities of small degree which do not fall into any of the series. Moreover, the book presents Arnold's classification algorithm which determines the singularity type of a given explicit equation. Notice that we did not try to find equations of the lowest possible degree for the singularities mentioned in the following challenge although there are certainly realizations of these singularities in lower degrees similar to the curves $f_{k,2,+}^2$ of degree $2k$ given in challenge 2 which define A_{2k^2-1}-singularities with a normal form of degree $2k^2$.

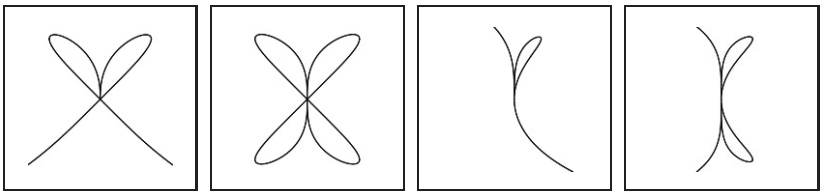


FIG. 19. *Some of the complicated singularities mentioned in challenge 15 for* $a = 5/7$.

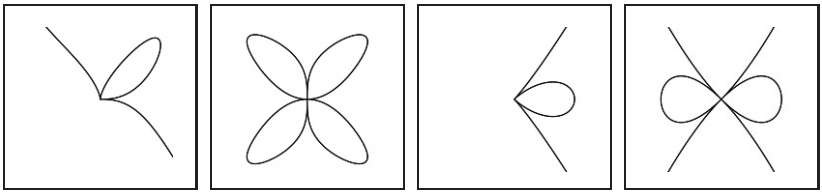


FIG. 20. *Some of the complicated singularities mentioned in challenge 15 for* $a = 5/7$.

CHALLENGE 15. *Visualize the plane curves*

name	*normal form*	*restrictions*
$J^{\pm}_{10+k}$	$x^3 \pm x^2y^2 + ay^{6+k}$	$a \neq 0,\ k > 0$
$X^{\pm\pm}_{9+k}$	$\pm x^4 + x^2y^2 \pm ay^{4+k}$	$a \neq 0,\ k > 0$
$Y^{\pm\pm}_{r,s}$	$\pm x^2y^2 \pm x^r + ay^s$	$a \neq 0,\ r, s > 4$
$\widetilde{Y}^{\pm\pm}_r$	$\pm(x^2 \pm y^2)^2 + ax^r$	$a \neq 0,\ r > 4$

for generic real values of the parameter a, e.g. $a = 5/7, 9/7, 13/7$.
Solution: We cannot give visualizations of all possible cases here. For a few images see Figures 19 and 20. Please consult our website [Lab03] for more pictures. Note that the number of halfbranches changes with the choice of the signs. E.g., $\widetilde{Y}^{++}_{4+2m}$ and X^{++}_{9+2m}, $m \in \mathbb{N}$, define solitary points although the number of complex branches is 4 in both cases. □

9. Other interesting examples. In this section we mention some examples which do not fit well into any of the other sections, but which are interesting from several points of view. We start with discriminants and then give some equations of plane curves on which several of the visualization issues discussed in previous sections occur at the same time.

9.1. Discriminants. The first examples we consider here are discriminants of polynomials in one variable which are of the form

$$f_{a,b}(x) = x^k + ax^l + b, \quad k > l > 0.$$

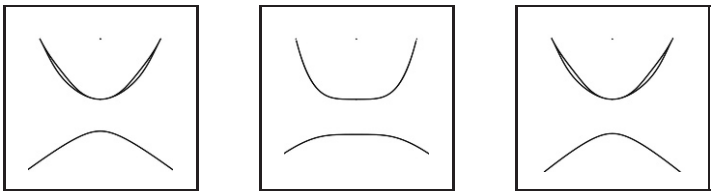


FIG. 21. *Some plane curves $SA_{k,l}$ with one higher solitary point and other singularities (see challenge 17).*

The ***discriminant*** D of $f_{a,b}$ is the polynomial which describes those parameters (a, b) for which $f_{a,b}$ has a double root, i.e. D is the resultant with respect to x of f and its derivative $f'(x)$. For $k = 4$, $l = 2$, this is obviously $b{\cdot}(4b - a^2)$ which has an A_2-singularity at $(a, b) = (0, 0)$. Most of the singularities occurring in this way have already appeared before, so that we do not make up a challenge for these.

Instead, we will have a look at a discriminant of two polynomials in two variables: Recently, while searching for trinomial systems with many real roots, the authors of [DRRS07] studied the discriminant of the system

$$x^6 + ay^3 - y,\ y^6 + bx^3 - x.$$

It turns out that this is the simplest known trinomial system to possess (a, b), e.g. $P = (44/31, 44/31)$, making a trinomial system have 5 roots in the positive quadrant (see [DRRS07] for proofs and for the equation of the discriminant). The nice thing about this example is the fact that a correct visualization of the plane curve D helped the authors to find the point P.

CHALLENGE 16. *Visualize the discriminant $D \in \mathbb{Q}[a, b]$ of the system*

$$x^6 + ay^3 - y,\ y^6 + bx^3 - x.$$

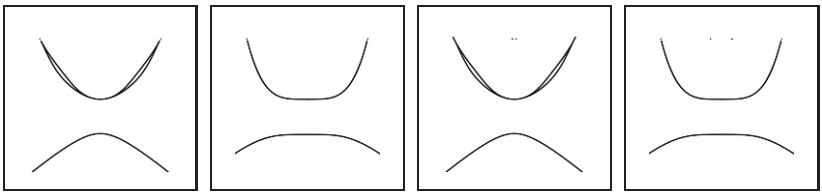
Solution: See [DRRS07] for many details on the curve. □

9.2. Plane curves with several difficulties. We now give some equations of plane curves which admit several of the challenges covered in the other sections at the same time. Our first examples have a high solitary point with a high zero–tangency and also singular points with high tangencies:

CHALLENGE 17. *Visualize the plane curves*

$$SA_{k,l}\colon (y - 1 - x^k)^l{\cdot}(y - x^k)^l + (y - 1)^{kl+1}y^{kl}$$

for $k, l = 2, 4, 6, \ldots$.
Solution: These curves have a solitary point in $(0, 1)$ and some other singularities with high tangencies. □



$SA_{2,4,0.001}$ $SA_{4,2,0.001}$ $SA_{2,4,-0.001}$ $SA_{4,2,-0.001}$

FIG. 22. *Some plane curves* $SA_{k,l,\varepsilon}$ *(see challenge 18).*

Now, we give some examples which have real singular points with high tangencies as well as several critical points with small critical values and singular points with imaginary coordinates:

CHALLENGE 18. *Visualize the plane curves*

$$SA_{k,l,\pm\varepsilon}\colon (y-1-x^k\pm\varepsilon)^l\cdot(y-x^k+\varepsilon)^l+(y-1)^{kl+1}y^{kl}$$

for $k,l=2,4,6,\ldots$ *and* $\varepsilon=10^{-m}$ *for* $m=1,2,3,\ldots$.

Solution: These curves have one singularity with high tangency at the origin. In addition, they have many critical points and many singularities, several of which have imaginary coordinates. □

To finish, here are some examples of plane curves which have a solitary point with a high tangency at the origin which is very close to a one-dimensional component of the curve. These will be very difficult to visualize for a numerical visualization software which treats small values as zero because for the examples given below the result will be incorrect, even topologically, in many cases (see Figure 23). A subdivision method which only has a heuristic stopping criterion for the recursive search will run into trouble for similar reasons.

CHALLENGE 19. *Visualize the plane curves*

$$SCA_{k,l,\pm\varepsilon}\colon \left((y-x^k)^l+y^{kl}\right)\cdot(y^2-x^2+\varepsilon)+y^{kl+2}$$

for $l=2,k=2,3,4,\ldots$ *and* $\varepsilon=10^{-m}$ *for* $m=1,2,3,\ldots$.

Solution: These plane curves of degree $2k+2$ have an $A^{\bullet}_{2k^2-1}$-singularity at the origin very close to which there is a one-dimensional real part of the plane curve (see Figure 24). □

A problem of a similar kind is the visualization of the following plane curves:

CHALLENGE 20. *Visualize the plane curves*

$$SAA_{k,l,\pm\varepsilon}\colon \left((y-x^k)^l+y^{kl}\right)\cdot((x-y^k)^l-x^{kl}-\varepsilon)+(xy)^{kl}$$

for $l=2,k=2,3,4,\ldots$ *and* $\varepsilon=10^{-m}$ *for* $m=1,2,3,\ldots$.

Solution: These plane curves of degree $4k$ have an $A^{\bullet}_{2k^2-1}$-singularity at the origin very close to which there is a one-dimensional real part of the plane curve (see Figure 25). □

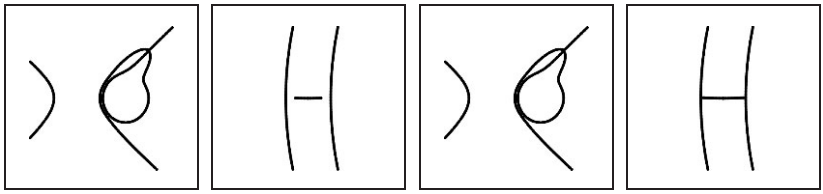


FIG. 23. *We visualize the plane curve $SAA_{2,2,0.001} + \tilde{\varepsilon}$ from challenge 20 for several values of $\tilde{\varepsilon}$ in order to visualize the difficulties which the curves $SAA_{2,2,0.001}$ pose to numerical software. The two leftmost pictures show the curve for $\tilde{\varepsilon} = 10^{-17}$ at two different zoom-levels; these pictures are topologically almost correct: the number of connected components is the right one, but the solitary point at the origin has been replaced by a small oval which looks like a short line. However, the two rightmost pictures ($\tilde{\varepsilon} = 10^{-16}$) are wrong, even topologically: locally, the three connected components become a single one.*

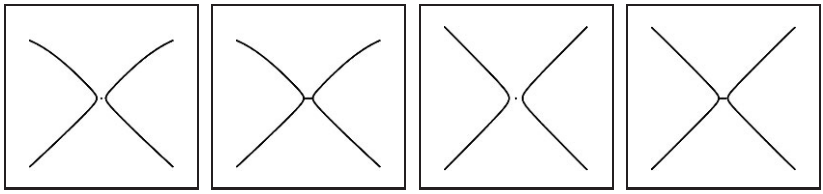


FIG. 24. *Some plane curves with a higher solitary point close to a one-dimensional component of the curve (see challenge 19). Algorithms treating small absolute values as zero produce topologically wrong results for such examples.*

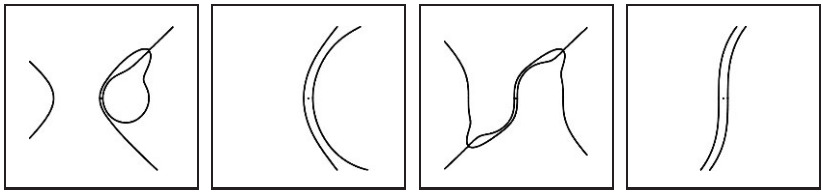


FIG. 25. *The plane curves $SAA_{2,2,0.001}$ and $SAA_{3,2,0.001}$, each at two different zoom-levels (see challenge 20).*

REFERENCES

[AGZV85] V.I. Arnold, S.M. Gusein-Zade, and A.N. Varchenko, *Singularities of differentiable maps*, Birkhäuser, 1985 (two volumes).

[Arn81] V.I. Arnold, *Singularity Theory*, London Math. Soc. Lecture Note Series, Vol. 53, Cambridge University Press, 1981.

[BCR98] J. Bochnak, M. Coste, and M.-F. Roy, *Real algebraic geometry*, Springer, 1998.

[BK86] E. Brieskorn and H. Knörrer, *Plane Algebraic Curves*, Birkhäuser, 1986.

[BR90] R. Benedetti and J.J. Risler, *Real algebraic and semi-algebraic sets*, Actualités Mathématiques. [Current Mathematical Topics], Hermann, Paris, 1990. MR 1070358 (91j:14045)

[Cos92] M. Coste, *Épaississement d'une hypersurface algébrique réelle*, Proc. Japan Acad. Ser. A Math. Sci. **68** (1992), no. 7, 175–180. MR 1193176 (94b:14056)

[Dim87] A. Dimca, *Topics on Real and Complex Singularities*, Vieweg, 1987.

[DRRS07] A. Dickenstein, J.M. Rojas, K. Rusek, and J. Shih, *Extremal Real Algebraic Geometry and A-Discriminants*, Moscow Mathematical Journal **7** (2007), no. 3, 425–452.

[Dur79] A.H. Durfee, *Fifteen characterizations of rational double points and simple critical points*, Enseign. Math.,II. Sér. **25** (1979), 132–163.

[GPS06] G.-M. Greuel, G. Pfister, and H. Schönemann, Singular *3.0*, A Computer Algebra System for Polynomial Computations, Centre for Computer Algebra, Univ. Kaiserslautern, 2006, `http://www.singular.uni-kl.de`.

[GZN00] S.M. Gusein-Zade and N.N. Nekhoroshev, *On singularities of type A_k on simple curves of fixed degree*, Funct. Anal. Appl. **34** (2000), 214–215.

[Har76] A. Harnack, *Ueber die Vieltheiligkeit der ebenen algebraischen Curven*, Math. Ann. **10** (1876), no. 2, 189–198. MR 1509883

[KO06] R. Kenyon and A. Okounkov, *Planar dimers and Harnack curves*, Duke Math. J. **131** (2006), no. 3, 499–524. MR 2219249

[Lab03] O. Labs, *Algebraic Surface Homepage. Information, Images and Tools on Algebraic Surfaces*, `www.AlgebraicSurface.net`, 2003.

[Shu98] E. Shustin, *Gluing of singular and critical points*, Topology **37** (1998), no. 1, 195–217.

[Var83] A.N. Varchenko, *On the Semicontinuity of the Spectrum and an Upper Bound for the Number of Singular Points of a Projective Hypersurface*, J. Soviet Math. **270** (1983), 735–739.

[Wes05] E. Westenberger, *Real hypersurfaces with many simple singularities*, Rev. Mat. Complut. **18** (2005), no. 2, 455–464.

A SUBDIVISION METHOD FOR ARRANGEMENT COMPUTATION OF SEMI-ALGEBRAIC CURVES

BERNARD MOURRAIN* AND JULIEN WINTZ*

Abstract. This chapter covers the use of subdivision methods in algebraic geometry with an emphasis on intersection, self-intersection, and arrangement computation, for the case of semi-algebraic curves with either implicit or parametric representation. Special care is given to the genericity of the subdivision, which can be specified whatever the context is, and then specialized to meet the algorithm requirements.

Key words. Symbolic-numeric computation, topology, intersection, arrangement, polynomial solvers, mathematical software.

AMS(MOS) subject classifications. 68W30, 65D17.

1. Geometric processing on semi-algebraic sets. In geometric modeling, the representation of shapes is naturally based on semi-algebraic models such as B-Spline parameterizations or implicit equations. A geometric object is described by assembling pieces of these primitives. When several objects have to be manipulated, for instance, to perform Constructive Solid Geometry (CSG) operations (*e.g.*, intersection, union, difference of volumes), the topological structure of the resulting shape has to be determined, together with a geometric description of the different elements which constitute this structure. Arrangement computation enables solving such problems. The first task is to determine the topology of one geometric object, that is, to analyse how this object decomposes the ambient space into connected regions, and how the boundaries of these regions are linked. The next step is to consider several objects, either incrementally or all at once. The arrangement computation describes the topological structure induced by the union of all the geometric primitives. This decomposition is made of connected regions, along with adjacency relationship on their boundary.

The effective use of such a decomposition in geometric computations, requires the ability to efficiently perform geometric queries such as point locations. In the case of a huge number of geometric primitives, manipulating the whole set of objects could be a bottleneck for many algorithms. It is thus critical to have the capacity to filter the computation, for instance, by exploiting a hierarchical structure with different levels of details.

The representation of shapes in application domains such as Computer Aided Geometric Design (CAGD) is based on semi-algebraic models such as B-Spline parameterizations. This yields compact representations which are easy to handle, through so-called control points. Piecewise linear models are also heavily used in practice, but they represent an approximation

*GALAAD, INRIA, BP 93, 06902 Sophia-Antipolis, France.

I.Z. Emiris et al. (eds.), *Nonlinear Computational Geometry*, The IMA Volumes in Mathematics and its Applications 151, DOI 10.1007/978-1-4419-0999-2_7,

of the shape which requires storing large amount of data to increase the degree of accuracy. The numerical stability of the computation performed on these models is a critical issue, which needs to be handled carefully, especially near singularities. This is typically the case when performing geometric operations such as intersection, which are intensively used in arrangement computations. In particular, the arrangement computation should be able to efficiently and robustly deal with object representation known with uncertainties, such as intersection points.

Arrangements of geometric objects is a field of computational geometry which has been studied for years [1], initially with simple objects such as line segments [4], circular arcs and curves are still investigated [22, 14, 11, 18] and can be used for computing an arrangement of surfaces [20]. The current methods mainly use a sweep approach [4]. They focus on events, which are critical points for a projection direction. The events are sorted before a critical value and the order after this critical value is deduced from information at the critical points.

More recently, sweep-line algorithms for computing an arrangement of arbitrary algebraic curves have emerged [8], making use of resultants to compute roots when the sweep-line encounters an event, another alternative is eigenvalue methods [18]. In [5] the authors present another context for computing an arrangement of a set of curves defined on a continuous two-dimensional parametric surface, while sweeping the parameter space. Finally, we mention an attempt at computing elements of an arrangement of implicit curves using interval arithmetic in a subdivision process [15].

When using sweep methods, events are treated when the sweep line encounters points of interest, where a projection on a line becomes critical, reducing the dimension of the problem but increasing its computational difficulty (for instance by computing resultants). Moreover, the projection step onto a subspace of smaller dimension is systematically followed by a lifting operation to come back to the initial space. Most of the existing approaches rely on exact geometric computation models. When dealing with segments, this is not really an obstacle, but for general semi-algebraic objects, these operations are delicate from a numerical point of view, since we are working at critical values. They require the manipulation of algebraic numbers, and their complexity could be a problem with large bitsize input polynomials.

In this paper, our aim is to describe a new method to compute arrangements of semi-algebraic sets, which allows filtering techniques, providing different level-of-details for the representation of the arrangement, and which requires as an external tool, the isolation of intersection points. The algorithm follows a subdivision approach, which focuses on regularity criteria and regular regions. We avoid the analysis at critical values by enclosing the singular points in a domain from which the problem at hand can be solved.

Subdivision algorithms appear in various fields of algebraic geometry such as topology, intersection and self-intersection computation, solution of non-linear equations. Such methods are less sensitive to numerical instability, while using approximations of objects and of their intersection points. Their application to arrangement computation has emerged such as in [7] where interval arithmetic is used to classify cells in the subdivision process. Subdivision methods are also very efficient for isolating the roots of polynomial equations, which appear in geometric problems [21, 9, 13, 19]. They have also been extended for the approximation of one or two dimensional objects [2, 16, 17].

The new method that we describe in this paper for computing arrangements of semi-algebraic curves aims at exploiting the power of these methods to localize the zeroes of polynomials. It is combining a subdivision approach with known algorithmic geometry schemes, bridging a gap between these two research areas. By storing geometric information on the zero locus of the functions which define the algebraic curves in hierarchical structures, we provide an efficient way to localize intersection points of region boundaries. It prevents useless computation by stopping the subdivision as soon as the topology of the object is known in a cell of subdivision. Moreover, the framework can be naturally extended to compute an arrangement of semi-algebraic surfaces in dimension 3.

The chapter is organized as follows. In the next section, we describe the generic subdivision scheme and the combination of ingredients applied on regions. In Section 3, we describe more specifically the local computation performed on these regions. In Section 4, we describe the tools needed to handle the semi-algebraic models we are manipulating. Finally, we show some examples in Section 5 and we conclude.

2. Subdivision approach and criterion of regularity. In this section, we describe in more details the generic subdivision approach, its behavior being the same whatever the representation of input objects is, as long as some representation specialized routines are provided. These specialized functions are described in the last section.

2.1. Subdivision. The subdivision process decomposes the initial domain into sub-domains in such a way that the structure (or the topology) of the objects inside these sub-domains is uniquely determined from information computed on the boundary. For that purpose, we need to check the existence and unicity of some characteristic points inside these domains. The method exploits, as a main ingredient, solvers which isolate the real roots of polynomial equations. The only requirement that these external solvers must satisfy is to enclose distinct solutions into boxes which are disjoint one from the other.

The input of the algorithm is a sequence $\Sigma = \{o_1 \dots o_n\}$ of semi-algebraic curves in the plane and an initial box $B_0 \subset \mathbb{R}^2$. Here is the

general scheme that we follow in the static case, *i.e.* considering all objects at once:

Algorithm 2.1: A generic subdivision algorithm

Input: a list of objects $\mathcal{O}$ and a box $B_0 \subset \mathbb{R}^2$.
Output: a list of regions.
Create a quadtree $\mathcal{Q}$ and set its root to B_0;
Create a list of cells $\mathcal{C}$ and initialize it with $[B_0]$;
while $\mathcal{C} \neq \emptyset$ **do**
  $c =$ `pop`$(\mathcal{C})$;
  $o =$ `objects`$(\mathcal{O}, c)$;
  if `regular`(o, c) **then**
    $\mathcal{Q} \leftarrow$ `topology`(o, c) ;
  else
    $\mathcal{C} \leftarrow$ `subdivide`(o, c) ;
  end
end
return `fusion`$(\mathcal{Q})$;

A quadtree is initialized with the initial bounding box B_0. Its root is appended to a list of cells. The function `objects` returns the list o of objects which are considered as active in c (those that intersect the cell). The algorithm checks the list o of objects lying in c for regularity. If not regular, the cell is subdivided into smaller cells, building a hierarchy of cells in the quadtree. These cells, appended to the list of cells are then checked for regularity and subdivided further if not regular. Finally regions are computed in regular cells and stored in the corresponding leaves of the quadtree and merged traversing the tree from its leaves to its root to obtain the set of regions determining the arrangement. The following operations remain to be clarified:

regular: the specific operation which checks if regions in a cell of the subdivision can be computed from the intersection of the active objects with the boundary, *i.e.* if the objects in the list are *regular* in the cell.

subdivide: the operation which subdivides a cell into four children, saving computation effort.

topology: the specific operation which computes regions in a regular cell.

fusion: the operation which reconstructs the connected components in the global arrangement from the regions stored in each leaf node of the tree.

2.2. Regularity criterion. Basically, an object in a cell, is said to be regular if the topology of the regions is uniquely determined from the intersection points of this object with the boundary of the cell. We have chosen the configurations, shown in Figure 1, in which the active objects are said to be regular.

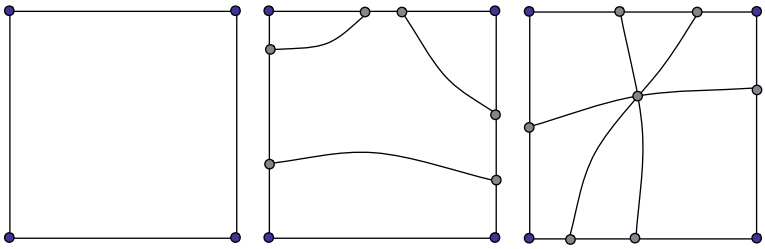


Fig. 1: *Regular cells.*

The first case (a) is the easiest configuration in which, the region is the whole cell. The object is then considered as non-active in the cell. The second case (b) is a cell in which the curve has either no x-critical point (with vertical tangent) or no y-critical point (with horizontal tangent). In this case, the topology of the curve inside the cell is uniquely determined from its intersection with the boundary. The connecting algorithm used to get the curve segments from points on the border of the cell will be described in Section 2.4. In the last case (c), all the branches of the object are intersecting at a unique (singular) point of the cell, in a star-shaped configuration. To analyse the topology around this point, we should be able to compute the number of branches stemming out from a self-intersection point. See [3] for the algorithm to obtain regions.

This set of conditions is sufficient to deduce the topological structure of each object in a cell as we will see in Sections 2.4 and 3.3. It involves the isolation of specific points of the curves and their insertion in x-regular or y-regular branches of these curves. This operation is supplied as a specialization of the generic subdivision arrangement algorithm. A specialisation of these functions for the different representations of semi-algebraic curves that we consider is described in Section 4.1.

This strategy of subdivision is to deal with some degenerate cases such as intersection points of more than 2 curves. It can also be optimised by requiring a limiting number of active objects or of branches of these objects in a cell. If, for instance, we require at most one branch per cell, the region computation will be simplified but the depth of the subdivision might increase, depending of the geometric configuration.

2.3. Subdividing a cell. When a cell is determined not regular, it is subdivided into smaller parts, generally more likely to be regular. When we subdivide the cell, we compute the intersection of the active objects in the cell with the new boundary and update the geometric information attached to the cell such as points of interest on the border or inside the cell.

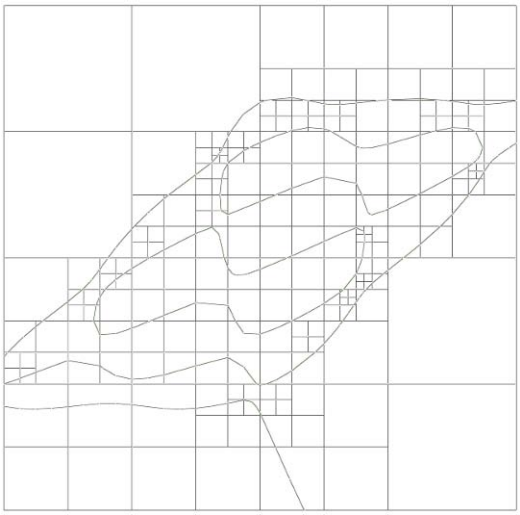


Fig. 2: *Subdivision algorithm.*

During the subdivision of a cell, points already computed will be distributed in child cells. We call this process vertical inheritance, since points of interest are distributed from a parent, to its children, in a tree hierarchy of cells. This inheritance is processed by locating the points in the children.

Another way to avoid computation is to consider adjacency relationships between child cells. Indeed, once a cell has been checked for regularity, generated points on the border of the cells can be inherited to its adjacent cells. This process takes place inside one level of the hierarchical tree, it is therefore called horizontal inheritance.

2.4. Topology. The topology step consists of computing the region structure defined by one curve o, from the information on the boundary of the cell. We assume that the curve is x or y-regular in the cell, that is with no vertical or no horizontal tangent. Suppose that we are in the first case of x-regularity. To each point p of o on the boundary of the cell, we associate an x-index defined as the sign of the scalar product of the normal to o at p interior to the cell and the unit vector in the x-direction. It is positive if a branch is appearing in the cell, when we sweep in the x-direction and negative if the branch is disappearing. If the sign is 0, depending on the multiplicity of intersection of the curve with the boundary and the orientation of the gradient, we duplicate or not the point and associate it with a positive and/or negative sign.

The algorithm of connection chooses a pair of consecutive points a, b on the boundary, so that a is of x-index $+$, b is of x-index $-$ and a has the smallest x-coordinate. The branch (a, b) is formed and the pair (a, b)

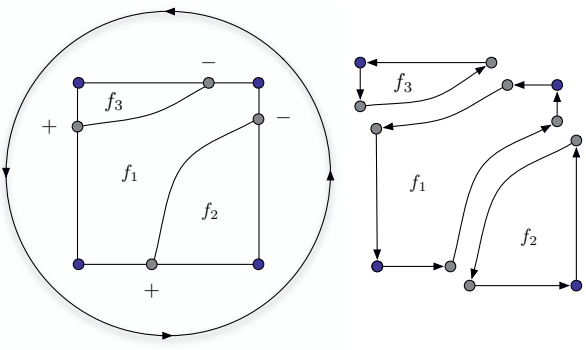


Fig. 3: *Generic scheme to compute a region from information on the border of a subdivision cell.*

is remove from the set of points on the boundary. This process is applied recursively, until no point is left on the boundary of the cell (see [3] for more details).

The construction of regions is based on this connection algorithm and illustrated in Figure 3. Regions are constructed while turning around the border of the cell and connecting border points together in loops, using the branch connections, as shown in Figure 3.

It is even more easy in the case of a star-shaped curve, since each intersection point of the object and the boundary of the cell is connected together with the singular point lying in the cell. See [3] for an algorithm to count the number of branches at a singular point.

When several objects are active in a cell and each of these objects is regular, regions defined by each object are computed as described above. They are then merged (or arranged), using the updating algorithm described in Section 3.3.

2.5. Fusion. At the end of the subdivision process, the quadtree contains regions computed in regular cells in its leaf nodes, internal nodes keep track of the subdivision structure. To get the set of regions defined by one object, or by a set of objects, omitting the subdivision process which locally ensures a correct topology in regular cells, these small regions have to be merged. To do so, we traverse the quadtree from its leaves to its root, merging the regions inside a level of the tree and across levels, in a process called *fusion*.

To be consistent regarding adjacency relationships, children nodes of an internal node are merged in the following order: the two cells located in top nodes, the two cells located in bottom nodes and the resulting top cell together with the resulting bottom cell. Regions determined in the

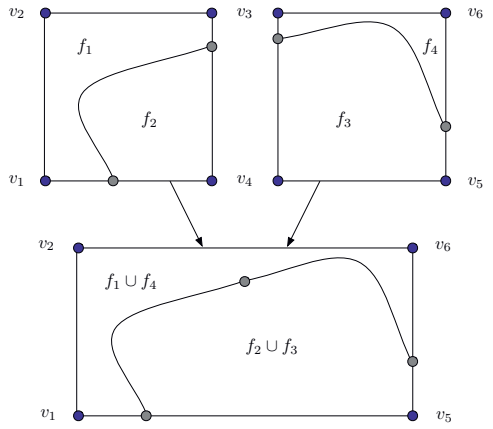


Fig. 4: *Fusion of regions.*

corresponding cells are merged together, resulting in a new set of regions which are the unions of adjacent regions as shown in Figure 4.

This algorithm brings local regions determined in regular cells associated with leaf nodes, up to the root, computing their union if they are adjacent across the levels of the tree. The root node finally contains regions determined by the object or the set of objects for which the subdivision process has been initiated.

Once these regions have been computed, they are inserted into an augmented influence graph, the data structure used to represent the decomposition of the space [6].

3. Region representation and computation. The subdivision scheme described above assures the transition from the (semi)algebraic representation of an object to its geometry, producing regions, either considering objects one by one or all at once.

This section focuses on defining in great details the structure of a region and the associated data used to facilitate the manipulation of regions.

3.1. Region representation. A region is traditionally defined by a set of elements of incremental dimensions with respect to the one of the input space: vertices, edges and faces. When computing an arrangement of semi-algebraic curves, regions are faces, incidence edges of which are curved segments with continuous representation. Vertices are some markers that help to somehow provide a discretization of these continuous objects.

The set of vertices of an arrangement is mainly defined by intersection operations on the input objects. In some cases, stronger conditions on the edges can be assumed such as a monotonicity requirement. In this case, x and y critical points will be inserted in order to turn the edges into mono-

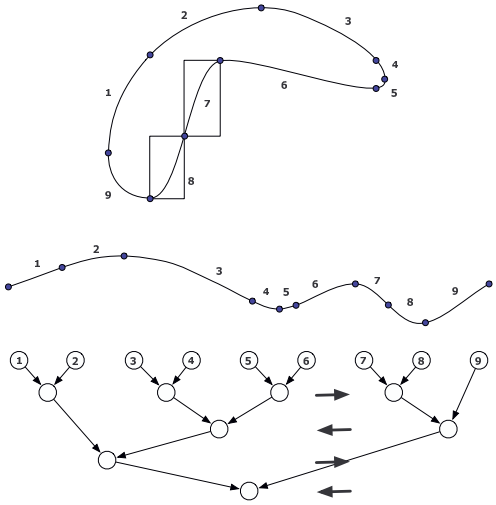


Fig. 5: *Building the region segmentation.*

tone ones. Also, some objects may have a representation which can induce degenerate points such as isolated singular points or self-intersections, usually referred to as singularities. These points appear explicitly as vertices in the arrangement, even if they are isolated. They provide a better accuracy in the description of a region and a certified topology.

The set of edges defining a region can also be constrained regarding some requirements. Indeed, we can afford the edges to be non intersecting, which means that a region will be "on the same side" of the oriented boundary. The orientation of edges is a paramount aspect when defining a region, so that Boolean operators can decide on which side of an edge lies the interior of the region.

This discretization is a main requirement for the input of usual algorithms dealing with regions such as Boolean operators, in which oriented edges are required to compute the resulting region.

This data structure can be enriched to enhance further operations in the case of a dynamic arrangement algorithm. Such algorithm maintains its solution under insertion or removal of objects, and conflict detection is a critical operation which needs to be performed efficiently.

3.2. Region segmentation. In the case where regions are computed for an object with no regard of existing objects, they may conflict with regions determined by another object, already inserted into the augmented influence graph.

To help locating such a conflict in efficient way, we associate with a region, an additional data structure called the *region segmentation*. It is

a binary tree in which leaf nodes are associated with an edge of a region, together with its corresponding bounding box, and internal nodes are associated with the bounding box of the union of their children bounding boxes. Figure 5 gives a hint on how to compute such a structure: the list of edges is flattened then traversed from left to right computing unions of boxes, in order to provide a balanced data structure.

Once regions have been processed to build the segmentation, it is very easy to find out whether two regions intersect, whatever the type of objects defining their edges are.

To do so, we "intersect" the respective segmentations associated with two regions, that is, beginning with the roots, we check the nodes and recursively proceed to their children as long as their associated boxes intersect. If two bounding boxes associated with two leaf nodes respectively (containing the edges) do intersect, the algorithm, shown below, provides a list of conflict zones, in which actual intersection points can be computed using representation specific procedures.

This query on the respective segmentations of two regions provides an efficient test to check regions for intersection and does not require any further algebraic computation. If objects do intersect, it will yield a list of reduced domains improving the efficiency of the algebraic tools used to isolate the intersection points.

Algorithm 3.1: Querying region segmentations for conflict

Input: two segmentation nodes n_1 and n_2
Output: a list $\mathcal{L}$ of conflict zones
if *!*`intersect(`n_1`,` n_2`)` **then**
  return $\mathcal{L}$;
end
if `isLeaf(`n_1`)` *and* `isLeaf(`n_2`)` **then**
  $\mathcal{L} \ll$ `intersect(`n_1`,` n_2`)` ;
else if `isLeaf(`n_1`)` *and* *!*`isLeaf(`n_2`)` **then**
  $\mathcal{L} \ll$ `query(`n_1`, left(`n_2`))` ;
  $\mathcal{L} \ll$ `query(`n_1`, right` *(*n_2*)*`)` ;
else if *!*`isLeaf(`n_1`)` *and* `isLeaf(`n_2`)` **then**
  $\mathcal{L} \ll$ `query(left(`n_1`),` n_2`)` ;
  $\mathcal{L} \ll$ `query(right(`n_1`),` n_2`)` ;
else
  $\mathcal{L} \ll$ `query(left(`n_1`), left(`n_2`))` ;
  $\mathcal{L} \ll$ `query(left(`n_1`), right(`n_2`))` ;
  $\mathcal{L} \ll$ `query(right(`n_1`), left(`n_2`))` ;
  $\mathcal{L} \ll$ `query(right(`n_1`), right(`n_2`))` ;
return $\mathcal{L}$;

3.3. Updating regions. This section addresses the problem of resolving conflicts between regions. This resolution consists of dividing con-

flicting regions into sets of regions, the union of which covers the conflicting regions. This set of regions is composed of the intersection of conflicting regions and of the difference of this intersection with original regions.

First, we consider the case of a static algorithm for computing an arrangement in a cell in which several objects are regular after the subdivision step. Secondly, we describe a dynamic algorithm which processes the regions defined incrementally by the objects. In this case, conflicts are dealt with directly during the subdivision process which takes into account intersections between the objects.

Let us describe first the static approach. When checking a cell containing several objects for regularity, we have to check whether pairs of objects do intersect or not. Let us assume that we have computed the regions $\mathcal{R}_0, \ldots, \mathcal{R}_s$ defined by the objects $o_1, \ldots, o_k$ in the cell. These are connected components the boundary of which consists of either sub intervals of the boundary of the cell or branches of the curves $o_1, \ldots, o_k$ connecting two points on the boundary of the cell.

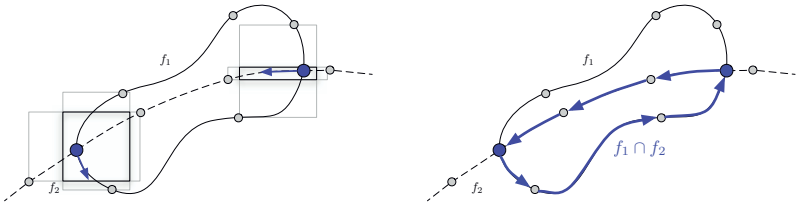
Consider now the regular object o_{k+1} and, for each of its branches, consider one of its end points on the boundary of the cell. It belongs to a region $\mathcal{R}_{i_0}$. We check where the object o_{k+1} intersects the objects on the boundary of $\mathcal{R}_{i_0}$.

If a branch of o_{k+1} does not intersect another object, we split the region $\mathcal{R}_{i_0}$ in two sub regions sharing the branch of o_{k+1} on their boundary.

If o_{k+1} does intersect another object o_{j_0} of the boundary of $\mathcal{R}_{i_0}$:

1. We insert the point on the corresponding branch of o_{j_0} and split the region $\mathcal{R}_{i_0}$ into two sub regions (deduced from the intersection point and the points of o_{k+1} on the boundary).
2. We take the region $\mathcal{R}_{i'_0} \neq \mathcal{R}_{i_0}$ which is sharing the branch of o_{j_0} on its boundary.
3. We repeat the intersection computation between o_{k+1} and the object on the boundary of the new region $\mathcal{R}_{i'_0}$.
4. If new points of intersection are found, the cell is subdivided and the same process is applied on the sub cells.
5. Otherwise the region $\mathcal{R}_{i'_0}$ is split in two sub regions (deduced from the intersection point and the points of o_{k+1} on the boundary).

Secondly, considering the dynamic algorithm, we process objects one by one, computing regions defined by an object in a subdivision process, and resolving conflicts with other regions in an efficient manner involving the region segmentation structure introduced here before, in a second time. The approach also allows us to directly compute Boolean operations on the regions defined by different objects, without computing the whole arrangement structure (see Figure 6 for the case of an intersection computation of two conflicting regions). Again, this Boolean operation will work just the same way whatever the representation chosen for objects defining the edges of the region, as long as intersection methods are provided in a specialization of it.



(a) Computing navigation information. (b) Walk-about border of regions.

Fig. 6: *A generic boolean operation.*

The idea behind this method, inspired from an algorithm proposed in [23] is to perform a walk-about on the border of regions, from an intersection point to another, regarding precomputed navigation information depending on the Boolean operation at hand. This method assumes the regions to be oriented in a way depending on the operation at hand. For the case of an intersection computation, edges have to be oriented counterclockwise (the interior then lies on the left of the edges), for the case of an union computation, edges have to be oriented clockwise (in which case the interior of the region lies on the right of the edges). This precondition can easily be met directly when computing the topology of regions inside regular cells (see Section 2.4).

First, using representation specific methods, we compute intersection points in the list of conflict zones provided by the query on the respective region segmentations. Then, to each one of those, we associate a so-called navigation information which decides on which region the walk-about continues, from an intersection point to another. This information can be computed in various ways also depending on the Boolean operation at hand. In the case of an intersection computation, the walk-about will continue on the region which edge is on the left of the other, in the case of a union computation, the walk-about will continue on the region, incident edge of which is on the right of the other (see Figure 6(a)). The resulting region is then constructed during the walk-about adding vertices as well as edges, considering adequate adjacency relationships (see Figure 6(b)).

4. Algebraic operation on semi-algebraic sets. We have presented a generic algorithm. Generic means that its overall behavior will always be the same, *assuming* some required functionalities on the specific object types . The algorithm therefore provides an abstraction of representation related functions. A generic algorithm can never be used out of the box, without providing these specific functionalities, it has to be *specialized*.

In the semi-algebraic curve arrangement algorithm example, these specific methods are mainly used for the regularity test and for the region intersection. Geometrically, they include the ability to compute:

- x and y critical points,
- self-intersection points or isolated points,
- cell-curve intersection points,
- curve-curve intersection points.

The only challenge left is to provide algebraic algorithms to compute these features. Having such tools, being able to express the underlying algebraic problems ends in being able to test the cells for regularity since we only have to count the number of features in a cell.

4.1. Isolating real roots. A critical operation, which we have to perform in an arrangement computation, is to isolate the roots of polynomial equations. In such a computation, we start with input polynomial equations (possibly with some incertitude on the coefficient) and we want to compute an approximation of the real roots of these equations or boxes containing these roots. Such operation should be performed very efficiently and with guarantee, since they are used intensively in geometric computations.

Hereafter, we will describe subdivision solvers which exploit the properties of *Bernstein bases*. The Bernstein basis is often used in a subdivision process involving algebraic curves since it is a very convenient representation to encode the restriction of a multivariate polynomial, within a given domain.

Univariate Bernstein basis. Given an arbitrary univariate polynomial function $f(x) \in \mathbb{K}$, we can convert it to a representation of degree d in Bernstein basis, which is defined by:

$$f(x) = \sum_i b_i B_i^d(x) \qquad \text{with } B_i^d(x) = \begin{pmatrix} d \\ i \end{pmatrix} x^i (1-x)^{d-i},$$

where b_i are usually referred to as control coefficients. Such a conversion is done through a basis conversion [10]. The above formula can be generalized to an arbitrary interval $[a, b]$ by a variable substitution $x' = (b-a)x + a$. We denote by $B_d^i = (x; a, b)\binom{d}{i}(x-a)^i(b-x)^{d-i}(b-a)^{-d}$ the corresponding Bernstein basis on $[a, b]$.

Using De Casteljau subdivision, the representation of a univariate polynomial in the Bernstein basis associated with any (sub-)interval can be deduced from its representation on the initial interval.

Multivariate Bernstein basis. The univariate Bernstein basis representation can be generalized to multivariate ones. Briefly speaking, we can rewrite the definition (see equation 4.1) in the form of tensor products. Suppose for $\mathbf{x} = (x_1, ..., x_n) \in \mathbb{R}^n$, $f = (\mathbf{x}) \in \mathbb{K}[\mathbf{x}]$ having the maximum degree $\mathbf{d} = (d_1, ..., d_n)$ has the form:

$$f(\mathbf{x}) = \sum_{k_1=0}^{d_1} ... \sum_{k_n=0}^{d_n} b_{k_1 ... k_n} B_{k_1}^{d_1}(x_1) ... B_{k_n}^{d_n}(x_n).$$

The De Casteljau subdivision for the multivariate case proceeds similarly to the univariate one, since the subdivision can be done independently with regards to a particular variable x_i. The Descartes' law also applies for the multivariate case. For a polynomial of n variables, the coefficients can be viewed as a tensor of dimension n.

Univariate Bernstein solver. Let us consider first an exact polynomial $f = \sum_{i=0}^{d} a_i x^i \in \mathbb{Q}[x]$. Our objective is to isolate the real roots of f, *i.e.* to compute intervals with rational endpoints that contain one and only one root of f, as well as the multiplicity of every real root. Here is the general scheme of the subdivision solver that we consider, augmented appropriately so that it also outputs the multiplicities. It uses an external function $V(f, I)$, which bounds the number of roots of f in the interval I.

Algorithm 4.1: Real root isolation

Input: A polynomial $f \in \mathbb{Q}[x]$, such that $\deg(f) = d$ and $\mathcal{L}(f) = \tau$.
Output: A list of intervals with rational endpoints, which contain one and only one real root of f and the multiplicity of every real root.

Compute the square-free part of f, *i.e.* f_{red} ;
Compute an interval $I_0 := (-B, B)$;
Initialize a queue Q with I_0 ;
while $Q \neq \emptyset$ **do**
  Pop an interval I from Q and compute $v := V(f, I)$;
  if $v = 0$ **then** discard I ;
  if $v = 1$ **then** output I ;
  if $v \geq 2$ **then** split I into I_L and I_R and push them into Q ;
end
Determine multiplicities of real roots, using f_{red} ;

Another interesting property of the univariate Bernstein representation related to Descartes rule of signs is that there is a simple yet efficient test for the existence of real roots in a given interval. It is based on the number of sign variation $V(\mathbf{b}_k)$ of the sequence $\mathbf{b}_k = [b_1, \ldots, b_d]$ that we define recursively as follows:

$$V(\mathbf{b}_k) = V(\mathbf{b}_{k-1}) + \begin{cases} 1 & \text{if } b_k b_{k-1} < 0 \\ 0 & \text{otherwise} \end{cases}$$

With this definition, we have:

PROPOSITION 4.1. *Given a polynomial* $f(x) = \sum_i^n b_i B_i^d(x; a, b)$, *the number* N *of real roots of* f *on* $]a, b[$ *is less than or equal to* $V(\boldsymbol{b})$, *where* $\boldsymbol{b} = (b_i)_{i=1\ldots n}$ *and* $N \equiv V(\boldsymbol{b}) \mod 2$.

With this proposition,

- if $V(\mathbf{b}) = 0$, the number of real roots of f in $[a, b]$ is 0.
- if $V(\mathbf{b}) = 1$, the number of real roots of f in $[a, b]$ is 1.

The approach can also be extended to polynomials with interval coefficients, by counting 1 sign variation for a sign sub-sequence $(+, ?, -)$ or $(-, ?, +)$, 2 sign variations for a sign sub-sequence $(+, ?, +)$ or $(-, ?, -)$, 1 sign variation for a sign sub-sequence $(?, ?)$, where $?$ is the sign of an interval containing 0. Again in this case, if a family $\overline{f}$ of polynomials is represented by the sequence of intervals $\bar{\mathbf{b}} = [\bar{b}_0, \ldots, \bar{b}_d]$ in the Bernstein basis of the interval $[a, b]$: if $V(\bar{\mathbf{b}}) = 0$, all the polynomials of the family $\overline{f}$ have no roots in $[a, b]$, if $V(\bar{\mathbf{b}}) = 1$, all the polynomials of the family $\overline{f}$ have one root in $[a, b]$.

This subdivision algorithm, using interval arithmetic, yields either intervals of size smaller than ϵ, which might contain the roots of $f = 0$ in $[a, b]$ or isolating intervals for all the polynomials of the family defined by the interval coefficients.

Multivariate Bernstein solver. We consider here the problem of computing the solutions of a polynomial system

$$\begin{cases} f_1(x_1, \ldots, x_n) = 0 \\ \vdots \\ f_s(x_1, \ldots, x_n) = 0 \end{cases}$$

in a box $B := [a_1, b_1] \times \cdots \times [a_n, b_n] \subset \mathbb{R}^n$.

The method for approximating the real roots of this system, that we describe now uses the representation of multivariate polynomials in Bernstein basis, analysis of sign variations and univariate solvers. The output is a set of small-enough boxes, which contain these roots. This subdivision solver can be seen as an improvement of the *Interval Projected Polyhedron* algorithm [21].

In the following, we use the Bernstein basis representation of a multivariate polynomial f of the domain $I := [a_1, b_1] \times \cdots \times [a_n, b_n] \subset \mathbb{R}^n$:

$$f(x_1, \ldots, x_n) = \sum_{i_1=0}^{d_1} \cdots \sum_{i_n=0}^{d_n} b_{i_1,\ldots,i_n} \, B_{d_1}^{i_1}(x_1; a_1, b_1) \cdots B_{d_n}^{i_n} x(x_n; a_n, b_n).$$

Definition 4.1. *For any $f \in \mathbb{R}[\mathbf{x}]$ and $j = 1, \ldots, n$, let*

$$m_j(f; x_j) = \sum_{i_j=0}^{d_j} \min_{\{0 \le i_k \le d_k, k \ne j\}} b_{i_1,\ldots,i_n} \, B_{d_j}^{i_j}(x_j; a_j, b_j)$$

$$M_j(f; x_j) = \sum_{i_j=0}^{d_j} \max_{\{0 \le i_k \le d_k, k \ne j\}} b_{i_1,\ldots,i_n} x \, B_{d_j}^{i_j}(x_j; a_j, b_j).$$

THEOREM 4.1 (Projection Lemma). *For any* $\mathbf{u} = (u_1, \ldots, u_n) \in I$, *and any* $j = 1, \ldots, n$, *we have*

$$m(f; u_j) \leq f(\mathbf{u}) \leq M(f; u_j).$$

As a direct consequence, we obtain the following corollary:

COROLLARY 4.1. *For any root* $\mathbf{u} = (u_1, \ldots, u_n)$ *of the equation* $f(\mathbf{x}) = 0$ *in the domain* I, *we have* $\underline{\mu}_j \leq u_j \leq \overline{\mu}_j$ *where*

- $\underline{\mu}_j$ *(resp.* $\overline{\mu}_j$*) is either a root of* $m_j(f; x_j) = 0$ *or* $M_j(f; x_j) = 0$ *in* $[a_j, b_j]$ *or* a_j *(resp.* b_j*) if* $m_j(f; x_j) = 0$ *(resp.* $M_j(f; x_j) = 0$*) has no root on* $[a_j, b_j]$,
- $m_j(f; u) \leq 0 \leq M_j(f; u)$ *on* $[\underline{\mu}_j, \overline{\mu}_j]$.

The solver proceeds in the following main steps:

1. applying a preconditioning step to the equations;
2. reducing the domain;
3. if the reduction ratio is too small, then split the domain;

until the size of the domain is smaller than a given ϵ.

The algorithm is parameterized by the preconditioning strategy, the reduction strategy and the subdivision strategy. It can be proved that the reduction based on the polynomial bounds m and M behaves like Newton iteration near a simple root, that is we have a quadratic convergence, using a local preconditioning. For more details on this method, including complexity bounds and practical behavior, see [19]. This approach is compatible with the sleeve techniques used for univariate polynomials, Using machine precision arithmetic, the guarantee that the computed intervals contain the roots of f, is obtained by controlling the rounding mode during the De Casteljau computation.

4.2. Implicit curves. In this section, we describe the two operations of the arrangement computation which are specialized for implicit curves, namely the isolation of specific points and the insertion of points on the branches of a regular curve.

We denote by $f_k(x, y) \in \mathbb{R}[x, y]$ the polynomial defining the implicit curve corresponding to the object o_k. The root isolation operations will be performed on the Bernstein representation of f_k on $D = [a, b] \times [c, d]$:

$$f_k(x, y) = \sum_{i=0}^{d_{x,k}} \sum_{j=0}^{d_{y,k}} b_{i,j}^k \, B_{d_{x,k}}^i(x; a, b) \, B_{d_{y,k}}^j(y; c, d),$$

where $B_d^i(x; u, v) = \binom{d}{i} (x - u)^i (v - x)^{d-i} \, (v - u)^{-d}$ (for $0 \leq i \leq d$, $u < v$). Here is the list of points that we have to compute:

• The intersection locus of an implicit curve and borders of a subdivision cell (*e.g.*, with coordinates $[x_{min}, x_{max}] \times [y_{min}, y_{max}]$) can be obtained by solving the univariate equations $f(x_{min}, y) = 0$ and $f(x_{max}, y) = 0$ in y for left and right vertical borders, and $f(x, y_{min})$ in x and $f(x, y_{max})$ in x for bottom and top horizontal borders.

• The critical points in the x-direction (resp. the y-direction) can be obtained by (resp.) solving:

$$\left\{ \begin{array}{rcl} f_k(x,y) & = & 0 \\ \partial_y f_k(x,y) & = & 0 \end{array} \right. \text{ and } \left\{ \begin{array}{rcl} f_k(x,y) & = & 0 \\ \partial_x f_k(x,y) & = & 0 \end{array} \right.$$

• To compute singular points (isolated or self-intersection points), we solve (*e.g.*, using the multivariate subdivision solver):

$$\left\{ \begin{array}{rcl} f_k(x,y) & = & 0 \\ \partial_x f_k(x,y) & = & 0 \\ \partial_y f_k(x,y) & = & 0 \end{array} \right.$$

• The intersection points of two curves o_k, o_l in a cell are obtained by solving the bivariate system:

$$\left\{ \begin{array}{rcl} f_k(x,y) & = & 0 \\ f_l(x,y) & = & 0 \end{array} \right.$$

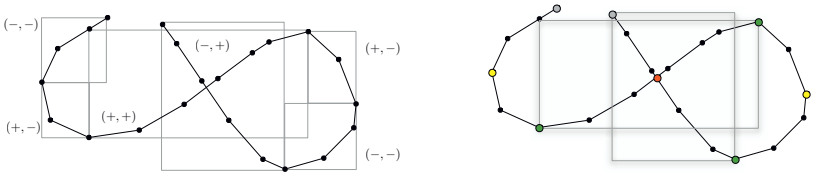
Finally, the insertion of a point p on one of the branches of a regular curve, say a x-regular curve is performed as follows: The point p is given by an isolating box $B = I_1 \times I_2 \subset \mathbb{R}^2$. We consider the x-monotone branches of $\mathcal{C}$, which overlap the interval $I_1 = [m_1 - \delta_1, m_1 + \delta_1]$ ($m_1 \in \mathbb{R}$, $\delta_1 > 0$). The localization of p is performed by computing the solutions $\alpha_1, \ldots, \alpha_{n_1}$ (resp. $\beta_1, \ldots, \beta_{n_2}$) of $f(m_1, y) = 0$ above (resp. below) the interval I_2. If $n_1 + n_2 + 1$ is not the total number of branches above the interval I_1, we refine the isolating box B. Otherwise, we the point p is on the $(n_1 + 1)^{\text{th}}$ branch of o above I_1 starting from the top.

4.3. Parametric curves. Parametrized objects feature many representations. We present the case of polynomial rational curves and the one of B-Spline curves, giving a hint on how to compute the elements needed for the specialization of our generic algorithm for these representations.

A uniform rational polynomial curve is defined by the formula:

$$c(t) = \left(\frac{x(t)}{w(t)}, \frac{y(t)}{w(t)} \right)$$

where $x(t), y(t), z(t)$ are polynomial functions evaluated to obtain the image of $t \in I \subset \mathbb{R}$ by c in $\mathbb{R}^2$.



(a) Segmenting the curve. (b) Querying the segmentation.

Fig. 7: *A generic procedure to compute features of a piecewise linear curves.*

It is possible to compute intersection points of a uniform rational curve with any line segment parallel to the axis, by solving the following univariate problems:

$$\begin{array}{lcll} y(t) - y_{min} w(t) & = & 0 & \text{and} \quad x_{min} \leq x(t) \leq x_{max} \\ y(t) - y_{max} w(t) & = & 0 & \text{and} \quad x_{min} \leq x(t) \leq x_{max} \\ x(t) - x_{min} w(t) & = & 0 & \text{and} \quad y_{min} \leq y(t) \leq y_{max} \\ x(t) - x_{max} w(t) & = & 0 & \text{and} \quad y_{min} \leq y(t) \leq y_{max} \end{array}$$

as well as the critical points:

$$x'(t)w(t) - x(t)\,w'(t) = 0 \quad y'(t)w(t) - y(t)w'(t) = 0$$

and self-intersection points: $c(t)|\exists(t,s), t \neq s$ verifying

$$\left\{ \begin{array}{lcl} \frac{x(t)w(s) - x(s)w(t)}{t-s} & = & 0 \\ \frac{y(t)w(s) - y(s)w(t)}{t-s} & = & 0. \end{array} \right.$$

The approach naturally extends to B-Spline curves defined as piecewise rational parameterized curves [10]. Since Descartes rule of sign is still valid for the representation of a curve c:

$$c(t) = \frac{\sum_{i=0}^{n} \mathbf{p_i} B_{i,d,\tau}(t)}{\sum_{i=0}^{n} w_i B_{i,d,\tau}(t)}$$

in the the B-Spline basis $B_{i,d,\tau}$ associated with the node sequence $\tau = (\tau_1, \tau_2, \ldots, \tau_{n+k})$ in degree d, the subdivision solvers described in Section 4.1 also work for this type of representation.

4.4. Piecewise linear curves. The treatment of piecewise linear curves $p_0, \ldots, p_s$ (with $p_i \in \mathbb{R}^2$) is similar to the case of parametric curves and is illustrated in Figure 7. First, to each segment p_i, p_{i+1} we associate a so-called "monotony code" (corresponding to the sign of the coordinates of the vector $\overrightarrow{p_i p_{i+1}}$. Then, segments are gathered in a *monotonous segmentation*, a balanced binary tree data structure, similar

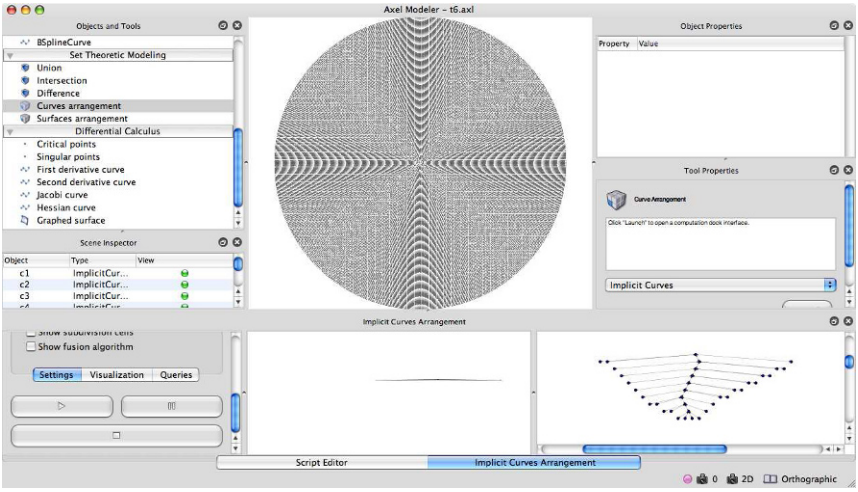


Fig. 8: *Computing an arrangement of 200 concentric circles in Axel.*

to the region segmentation which is queried the same way to compute self-intersection points. The only difference between the region segmentation and the monotonous segmentation is that in the latter, curve segments are gathered with regard of their monotony code, $M_k = M(t_k, t_{k+1}) = (sign(x(t_{k+1}) - x(t_k)), sign(y(t_{k+1}) - y(t_k)))$, Figure 7(a). These segmentations are then compared to find non adjacent intersecting nodes (comparing the coordinates of their associated edges' bounding boxes) in which, intersection points are computed using usual methods 7(b).

5. Examples. Rather than analysing the complexity of the algorithm (which is not a trivial task), this section explains how it behaves on some examples. Since piecewise linear and parametric representations do not represent a high computational challenge, we will focus on implicit representation.

The algorithm has been implemented within the algebraic geometric modeler Axel [1]. The software is a "proof of concept" prototype and focuses on genericity and if no optimization is provided, the algorithm is slowed down for visualization purposes. Timings are only given as an indication of the practical behavior. Experiments have been run on an Intel 4 CPU 3.40GHz computer with 1024Mo. RAM.

5.1. Concentric circles. The first example, as shown in Figure 8 is the one of 200 concentric circles given implicitly by the equation $x^2 + y^2 - r^2 = 0$ where r is the radius of the circle, considered from 1 to 200.

[1] `http://axel.inria.fr.`

This example illustrates the importance of the choice of a set of regularity criteria used to drive the subdivision process and the topology computation of regions. Indeed, with a large amount of curves, close one to the other, considering only one segment of curve in a regular cell would lead to a huge number of subdivisions. Increasing the depth of the quadtree and therefore producing a bigger set of local regions in regular cells inducing a longer fusion process. Considering the regularity criterion exposed in Section 2.2 and the connecting algorithm explained in Section 2.4, the algorithm behaves as follows:

Subdivision process: 11.919s
- # curve-cell intersections: 1240
- # curve-curve intersections: 0
- # singular points: 0
- # critical points: 800
- quadtree depth: 8

Fusion process: 42.465s
- # regions: 201.

5.2. Singular implicit curves in degenerate configuration. In the next example, we consider a set of 15 implicit curves in degenerate configuration. Among the curves, three of them are singular, including a Descartes pholium and two other singular curves, one which contains 4 intersecting circles at the singular point: $x^8+4x^6y^2+6y^4x^4+4y^6x^2+y^8-4x^6-12y^2x^4-12y^4x^2-4y^6+16x^2y^2$, the other one with 4 branches at another coincident singular point: $x^6+y^2x^4-y^4*x^2-2x^4-y^6+2y^4+x^2-y^2+xy$, see Figure 9.

To prevent large subdivision at the intersection locus of the two previous curves we use the topological degree [3] of the curve to compute the number of branches stemming out from the singularity, and, compared to the number of intersection points on the border of the cell we can detect a star shape structure and compute the topology accordingly.

Subdivision process: 7.224s
- # curve-cell intersections: 865
- # curve-curve intersections: 112
- # singular points: 7
- # critical points: 63
- quadtree depth: 10

Fusion process: 15.675s
- # regions: 142.

This last example has been computed with a reduced set of regularity criteria using univariate subdivision solvers to produce curve-cell intersection points while curve-curve intersection points, critical and singular points are computed by the mean of multivariate subdivision solvers.

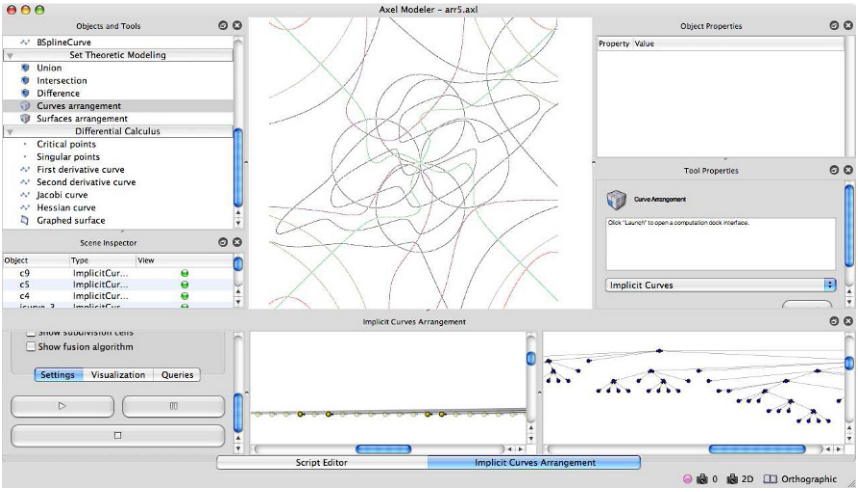


Fig. 9: *Computing an arrangement of 15 implicit curves of degree up to 8 in degenerate configuration in Axel.*

More examples and illustrations of the algorithm can be found on the Axel's website [2] [3].

Summary and outlook. In designing the arrangement algorithm, special attention has been given to genericity, combining existing subdivision algorithms in various fields of algebraic geometry such as topology, intersection and self-intersection computation, together with known algorithmic geometry schemes, bridging a gap between these two research areas.

Another emphasis has been put on exactness by providing certified methods. For example, using topological degree, we are able to ensure a geometric configuration even though we use approximate tools for obtaining the points. This certification relies directly on the isolation certification provided by the external polynomial solver.

Finally, this algorithm is currently partially implemented in the algebraic geometric modeler Axel, making extensive use of design patterns [12] to be consistent with its generic design. A special care has been given to keeping underlying data structures accessible to the user for queries such as point locations or Boolean operations.

We are currently investigating the adaptation of the method to higher dimension for computing an arrangement of semi-algebraic surfaces. Since the subdivision scheme naturally extends to any dimension, this extension only requires the provision of specialized implementation of algebraic operations.

[2] `http://axel.inria.fr/user/screenshots`.
[3] `http://axel.inria.fr/user/screencasts`.

REFERENCES

[1] P. Agarwal and M. Sharir, *Arrangements and their applications*, Handbook of Computational Geometry (J. Sack, ed.) (2000), pp. 49–119.

[2] L. Alberti, G. Comte, and B. Mourrain, *Meshing implicit algebraic surfaces: the smooth case*, in Mathematical Methods for Curves and Surfaces: Tromso'04, L.S.M. Maehlen, K. Morken, ed., Nashboro, 2005, pp. 11–26.

[3] L. Alberti and B. Mourrain, *Visualisation of implicit algebraic curves*, in Pacific Graphics, M. Alexa, S. Gortler, and T. Ju, eds., Lahaina, Maui, Hawai, USA, 2007, IEEE Computer Society, pp. 303–312.

[4] J.L. Bentley and T.A. Ottmann, *Algorithms for reporting and counting geometric intersections*, IEEE Trans. Comput., C-28 (1979), pp. 643–647.

[5] E. Berberich, E. Fogel, D. Halperin, K. Mehlhorn, and R. Wein, *Sweeping and maintaining two-dimensional arrangements on surfaces: A first step*, in ESA, 2007, pp. 645–656.

[6] J.-D. Boissonnat and M. Yvinec, *Algorithmic geometry*, Cambridge University Press, New York, NY, USA, 1998.

[7] A. Bowyer, J. Berchtold, D. Eisenthal, I. Voiculescu, and K. Wise, *Interval methods in geometric modeling*, Geometric Modeling and Processing (2000), pp. 321–327.

[8] A. Eigenwillig and M. Kerber, *Exact and efficient 2d-arrangements of arbitrary algebraic curves*, in Proc. of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms, accepted for publication, 2008.

[9] G. Elber and M.S. Kim, *Geometric constraint solver using multivariate rational spline functions*, in Proceedings of the sixth ACM Symposium on Solid Modelling and Applications, ACM Press, 2001, pp. 1–10.

[10] G. Farin, *Curves and surfaces for computer aided geometric design: a practical guide*, Comp. Science and Sci. Computing, Acad. Press, 1990.

[11] E. Fogel, D. Halperin, L. Kettner, M. Teillaud, R. Wein, and N. Wolpert, *Arrangements*, in Effective Computational Geometry for Curves and Surfaces, J.-D. Boissonnat and M. Teillaud, eds., Springer-Verlag, Mathematics and Visualization, 2006, pp. 1–66.

[12] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, *Design patterns: elements of reusable object-oriented software*, Addison-Wesley Professional, 1995.

[13] J. Garloff and A.P. Smith, *Investigation of a subdivision based algorithm for solving systems of polynomial equations*, in Proceedings of the Third World Congress of Nonlinear Analysts, Part 1 (Catania, 2000), Vol. **47**, 2001, pp. 167–178.

[14] D. Halperin, *Arrangements*, in Handbook of Discrete and Computational Geometry, J. E. Goodman and J. O'Rourke, eds., CRC Press LLC, Boca Raton, FL, 2004, ch. 24, pp. 529–562.

[15] Y. Hijazi and T. Breuel, *Computing arrangements using subdivision and interval arithmetic*, in Proceedings of the Sixth International Conference on Curves and Surfaces Avignon, 2006, pp. 173–182.

[16] S. Joon-Kyung, E. Gershon, and K. Myung-Soo, *Contouring 1- and 2-Manifolds in Arbitrary Dimensions*, in SMI'05, 2005, pp. 218–227.

[17] C. Liang, B. Mourrain, and J.-P. Pavone, *Subdivision Methods for the Topology of 2d and 3d Implicit Curves*, in Geometric Modeling and Algebraic Geometry, Bert Juetller and Ragni Piene, eds., Springer, 2007, pp. 199–214.

[18] V. Milenkovic and E. Sacks, *An approximate arrangement algorithm for semi-algebraic curves*, in SCG '06: Proceedings of the twenty-second annual symposium on Computational geometry, New York, NY, USA, 2006, ACM Press, pp. 237–246.

[19] B. Mourrain and J.-P. Pavone, *Subdivision methods for solving polynomial equations*, J. of Symbolic Computation, doi:10.1016/j.jsc.2008.04.016 (2009), pp. 1–15. To appear (see also http://hal.inria.fr/inria-00070350/).

[20] B. Mourrain, J.-P. Técourt, and M. Teillaud, *On the computation of an arrangement of quadrics in 3d*, Comput. Geom. Theory Appl., **30** (2005), pp. 145–164. Special issue, 19th European Workshop on Computational Geometry, Bonn.

[21] E. Sherbrooke and N. Patrikalakis, *Computation of the solutions of nonlinear polynomials systems*, Computer Aided Geometric Design, **10** (1993), pp. 379–405.

[22] N. Wolpert, *Jacobi curves: Computing the exact topology of arrangements of non-singular algebraic curves*, In ESA 2003, LNCS 2832, **12** (2003), pp. 532–543.

[23] B. Zalik, M. Gombosi, and D. Podgorelec, *A quick intersection algorithm for arbitrary polygons*, in SCCG98 Conf. on Comput. Graphics and its Application, L.S. Kalos, ed., 1998, pp. 195–204.

INVARIANT-BASED CHARACTERIZATION OF THE RELATIVE POSITION OF TWO PROJECTIVE CONICS

SYLVAIN PETITJEAN*

Abstract. In this paper, we give predicates of bidegree at most $(6,6)$ in the input for characterizing the relative position of two projective conics. By relative position we mean the morphology of the intersection, the rigid isotopy class and which conic is inside the other when applicable. The predicates are derived by analyzing the algebraic invariant theory of pencils of conics and related constructions.

1. Introduction. Geometric computing with curved objects is often plagued with robustness issues. For instance, most commercial modeling or CAGD software choke on near-degenerate problem instances.

Various attempts at better handling degeneracies among non-trivial objects have led to the unfolding of the paradigm of *exact geometric computing*. Recall that a geometric object is really two things: a combinatorial structure (which for instance encodes the incidence properties of the elements constituting the object) and numerical quantities (coordinates) describing the embedding of the object in space. Since there are consistency constraints governing the relation between combinatorial information and numerical quantities, the numerical instability of geometric algorithms is intimately linked to this double nature of geometric objects. Working under the paradigm of exact geometric computing means doing calculations in which numerical quantities are evaluated to sufficient precision (exactly if needed) in order for the underlying combinatorial structure to be mathematically exact.

The dependence of combinatorial decisions on numerical computations is encapsulated in the notion of *geometric predicates*. Evaluating a geometric predicate often means determining the sign of a polynomial expression in the input coefficients. The sign itself encodes the answer to a simple geometric query like "are two given surfaces disjoint, tangent, or transversally intersecting?". The paradigm of exact geometric computing requires the predicates to be evaluated exactly, ensuring that the branchings of the algorithm are correct, that the software will not crash, loop indefinitely or output a wrong answer, and thus that the topological structure of the output is correct.

To improve algorithm performance, the evaluation of predicates is in general performed using a mix of interval, floating-point arithmetic (when the value of the expression is sufficiently far from zero) and exact arithmetic (otherwise). The degree of the polynomials expressing geometric predicates is a direct measure of the required number of bits and therefore of the efficiency of the implementation: the higher the degree of the predicate,

*LORIA-INRIA, Campus scientifique, BP 239, 54506 Vandœuvre-lès-Nancy cedex, France (Sylvain.Petitjean@loria.fr).

I.Z. Emiris et al. (eds.), *Nonlinear Computational Geometry*, The IMA Volumes in Mathematics and its Applications 151, DOI 10.1007/978-1-4419-0999-2_8,

the more often the arithmetic filters will fail and the more costly the exact evaluation will be. Translating each geometric decision in predicates of (as) low degree (as possible) is therefore critical to strictly limit the arithmetic demands of the implementation.

Past work. Finding simple predicates for determining the intersection type and relative position of a pair of simple curved objects has recently received a lot of attention. Wang and Krasauskas [31] and Liu and Chen [22] obtained the characterization of certain positions of ellipses using techniques from computer algebra, but gave no complete classification. Etayo *et al.* [12] have completely tackled the case of two ellipses using tools from real algebraic geometry (Sturm-Habicht sequences). Systematizing those ideas and adopting a more global point of view, Briand [4] characterized the rigid isotopy classes of two arbitrary conics. Some of the predicates obtained have however unnecessary high degrees, as we shall see.

A lot less is known in 3D. Wang *et al.* [32] have characterized what it means for two ellipsoids to be separated. More recently, Tu *et al.* [28] and Dupont *et al.* [10] have showed two somewhat different ways of extracting information from a pencil of quadrics so as to obtain exact morphological classification. However, they don't exhibit simple, discriminating polynomial functions of the input and control on the degree of the predicates is largely lost by intermediate constructions.

Contributions. In this paper, we consider the problem of characterizing the type of the intersection and the relative position of two plane projective conics with simple, low-degree predicates. More precisely, we prove the following:

THEOREM 1.1. *The real projective type of the intersection of a pair of conics of* $\mathbb{P}^2(\mathbb{R})$*, their rigid isotopy class and which conic is inside the other in nested cases can be determined with predicates of bidegree at most* $(6, 6)$ *in the coefficients of the conics.*

Instrumental in the proof of this result is the recourse to *algebraic invariant theory*, i.e. the study of the intrinsic properties of polynomial systems. Algebraic invariants were a hot topic in the 19th century. They were perceived as a bridge between geometry and algebra by the mathematicians of that era (culminating with Klein's famous Erlangen Program, and the view of geometry as the study of the properties of a space that are invariant under the action of a group of transformations).

It is only natural that algebraic invariants should appear in this context, since the type of the intersection of two conics is unchanged by projective transformations. However, their relevance to the fields of geometric computing and computational geometry has hitherto been largely unnoticed. People have occasionally stumbled upon quantities of invariantal nature (see for instance [8] on the sweeping of arrangements of circle arcs) but have usually failed to grasp their significance. For these reasons, we have chosen to give in this paper a fairly detailed introduction to this topic from a classical perspective.

The invariant theory of pairs of conics is an old subject that was explored by the classics (cf. [15]). That of pencils of conics, which is more relevant for studying properties of the intersections, is less well-known: it has been investigated by Todd [26, 27] around the mid of the 20th century, though only over the complex numbers. Briand [4] has also elements of the complex invariant theory of conics, but uses tools from real algebra (subresultant sequences) to conclude in the real case. Building upon Todd and Briand's work, we are able to fully exploit the invariant theory of pencils of conics to improve upon known results and produce a set of simple, low-degree predicates.

Paper outline. The organization of the paper is as follows. After some notations and preliminaries (Section 2), we enumerate in Section 3 the different orbits of pencils of conics under the action of the projective linear group and the classes of pairs of conics under rigid isotopy. Section 4 gives a gentle introduction to algebraic invariant theory and presents the main tools needed for computing invariants and covariants. The invariant theory of pairs of conics is developed in Section 5. Then we consider in Section 6 a special type of invariant objects, called combinants, which are intrinsically attached to pencils of conics. We put things together and show in Section 7 how to distinguish between pencil orbits with low-degree predicates. In the nested cases, we also show in Section 8 how to decide which conic is inside the other. Finally, we give some examples in Section 9, one of which involves a conic depending on a parameter, before concluding.

2. Preliminaries. In what follows $\mathbb{P}^k(\mathbb{K})$ denotes the projective space of dimension k over the field $\mathbb{K}$. A *conic* of the projective plane $\mathbb{P}^2(\mathbb{R})$ is two things:

- an algebraic object: an element of $\mathbb{P}(S_2(\mathbb{R}^{3^*}))$, the projectivization of the space $S_2(\mathbb{R}^{3^*})$ of real ternary quadratic forms. The algebraic conic associated to the quadratic form Q_S is denoted by $[Q_S]$;
- a geometric object: the zero set, in $\mathbb{P}^2(\mathbb{R})$, of the algebraic conic, denoted by $[Q_S = 0]$.

(We will distinguish between these different objects only when needed.) A conic is *proper* or *non-singular* if its quadratic form is; it is *degenerate* otherwise.

Recall that a real (resp. complex) *congruence* is a transformation of the general linear group $\mathrm{GL}_n(\mathbb{R})$ (resp. $\mathrm{GL}_n(\mathbb{C})$), represented by an invertible $n \times n$ matrix with entries in $\mathbb{R}$ (resp. $\mathbb{C}$). The subgroup of $\mathrm{GL}_n(\mathbb{K})$ consisting of matrices with determinant equal to 1 is the special linear group $\mathrm{SL}_n(\mathbb{K})$.

Attached to the quadratic form Q_S of a conic is a real 3×3 symmetric matrix S such that $Q_S = \mathbf{x}^T S \mathbf{x}$, where $\mathbf{x}$ are coordinates of $\mathbb{R}^3$. Note that a quadratic form Q_S is proper iff $\det S \neq 0$. Since S is symmetric, all of its eigenvalues are real. Let σ^+ and σ^- be the numbers of positive and negative eigenvalues of S, respectively. The *rank* of S (and Q_S) is the sum of σ^+ and

σ^-. The *inertia* of S (and Q_S) is the ordered pair (σ^+, σ^-). By Sylvester's Inertia Law, the inertia is invariant by real congruence transformations, i.e. $\forall P \in \mathrm{GL}_3(\mathbb{R})$ the matrices S and $P^T S P$ have the same inertia [20]. Actually, the inertia encodes all the invariant information of a real ternary quadratic form under the action of $\mathrm{GL}_3(\mathbb{R})$.

The *dual quadratic form* associated to a quadratic form Q_S of $\mathbb{R}^3$ is the quadratic form $Q_{S'}$ on $\mathbb{R}^{3^*}$ whose matrix S' is the adjoint of S, i.e. the transpose of the cofactor matrix, written $\mathrm{adj}\,(S)$. Note that if Q_S has rank 3 (resp. 2, 1, 0), then $Q_{S'}$ has rank 3 (resp. 1, 0, 0). The conic *dual* to $[Q_S]$ is $[Q_{S'}]$. The corresponding geometric conic is a subset of the dual projective space $\mathbb{P}^2(\mathbb{R})^*$ and represents the set of lines tangent to $[Q_S = 0]$.

A *pencil of conics* is a line in $\mathbb{P}(S_2(\mathbb{R}^{3^*}))$. The pencil is said to be *non-degenerate* if it contains proper conics. All conics of a given pencil share a set of common points, called the *base points* of the pencil. They are the intersection points of any two distinct conics of the pencil. In the complex projective space, non-degenerate pencils of conics always have four base points, when counted with multiplicities.

Given two ternary quadratic forms Q_S and Q_T, the line spanned by $[Q_S]$ and $[Q_T]$ is called the pencil generated by Q_S and Q_T. Attached to this pencil is a binary cubic

$$\mathcal{D} = \det(\lambda S + \mu T)$$

called the *characteristic form* of the pencil.

3. Classification of intersections of conics. There are nine orbits of non-degenerate pencils of conics under the action of $\mathrm{PGL}_3(\mathbb{R})$, i.e. the projective linear group. Their identification was obtained by Levy [21] (and also follows from Uhlig's canonical form of a pair of real symmetric matrices [30]). Table 1 gives representatives of each orbit and the morphology of the (real) intersection associated with each orbit. The top part of this table focusses on non-degenerate pencils, for which the orbits belong to five groups (denoted I, II, III, IV and IV using Levy's nomenclature) corresponding to distinct morphologies over the complex numbers. Group I corresponds to generic intersections of two conics, i.e. four simple (real or complex) points. For the sake of completeness, the bottom part of the table also lists representatives of orbits of degenerate pencils (the nomenclature is ours), stopping when the pencil of conics boils down to a pencil of binary quadratic forms.

The classification of ordered pairs of real projective conics modulo rigid isotopy, which corresponds to a real deformation of the equations of the conics that does not change the nature (real or complex, multiplicity) of the intersection points of the two conics, was obtained by Gudkov as a byproduct of his extensive classification of real curves of degree 4 in the real projective plane (see [16] for a brief survey in English as well as references to the original Russian articles). Briand [4] re-establishes this classification,

TABLE 1

Representatives of pencil orbits under the action of $\mathrm{PGL}_3(\mathbb{R})$ *and associated morphology.*

Orbit	**Real morphology**	Q_S	Q_T
I	four simple points	$x^2 - z^2$	$x^2 - y^2$
Ia	empty set	$x^2 + z^2$	$x^2 + y^2$
Ib	two simple points	yz	$x^2 + y^2 - z^2$
II	two simple points and a double point	$y^2 - z^2$	xy
IIa	a double point	$y^2 + z^2$	xy
III	two double points	z^2	$x^2 - y^2$
IIIa	empty set	$x^2 + y^2$	z^2
IV	a simple point and a triple point	$xz + y^2$	yz
V	a quadruple point	y^2	$z^2 + xy$
VI	conic	$x^2 + y^2 - z^2$	$x^2 + y^2 - z^2$
VIa	empty set	$x^2 + y^2 + z^2$	$x^2 + y^2 + z^2$
VII	$\mathcal{D} \equiv 0$, no singular point in common	xy	xz
VIII+	$\mathcal{D} \equiv 0$, common singular point	pencil of binary quadratics	

proving that each rigid isotopy class is determined by an orbit of pencils of conics under $\mathrm{PGL}_3(\mathbb{R})$ and the position of the two conics with respect to the singular conics in the pencil they generate. Call arc of the pencil a maximal range of pencil parameters not corresponding to a singular conic. Label a class by an N (for "neighbors") when its representatives are on a same arc of the pencil and by an S (for "separated") otherwise. Briand shows that there are 14 equivalence classes for *unordered* pairs of proper non-empty conics under rigid isotopy and exchange, of which Figure 1 gives a graphical representation.

Among the 14 classes for pairs,

$$\mathsf{IN}, \mathsf{IS}, \mathsf{IaS}, \mathsf{IbN}, \mathsf{IIS}, \mathsf{IIaS}, \mathsf{IIIS}, \mathsf{IVN}$$

are also equivalence classes for *ordered* pairs of proper non-empty conics under rigid isotopy. In addition, each class of pairs among

$$\mathsf{IaN}, \mathsf{IIaN}, \mathsf{IIIN}, \mathsf{IIIaN}, \mathsf{VN}$$

splits into two classes for ordered pairs, corresponding to one conic lying inside[1] the other. Finally, the class of pairs IIN splits into two classes for ordered pairs, corresponding to one conic lying inside the other in the neighborhood of the double point. Table 2 gives representatives for these last six classes.

[1] Note that it makes sense to talk about inside/outside since any proper non-empty conic cuts out the real projective plane into two connected components, one homeomorphic to a Möbius band – the outside – and the other homeomorphic to an open disk – the inside.

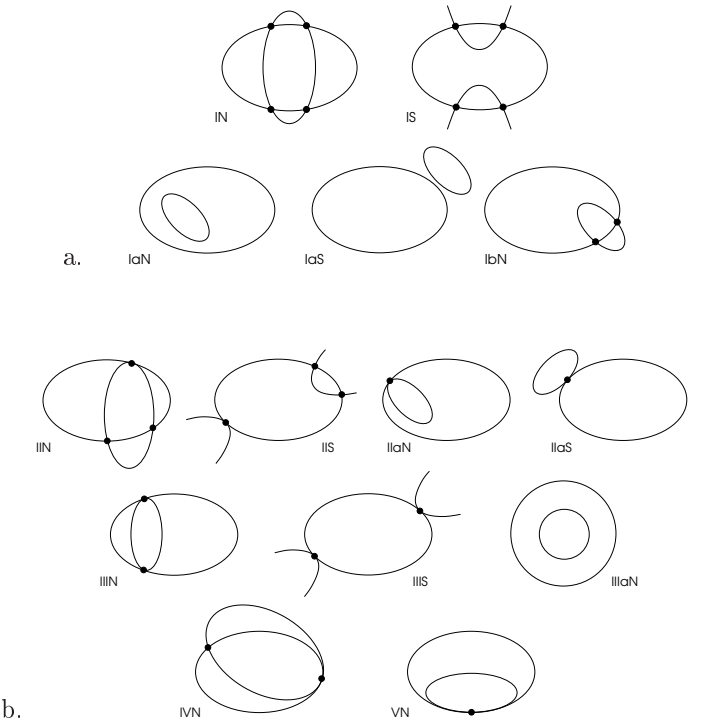


FIG. 1. *Morphology of the intersection for a pair of representative conics inside each rigid isotopy class. a. Generic cases. b. Non-generic cases.*

TABLE 2

Representatives for equivalence classes of pairs of proper non-empty conics under rigid isotopy leading to two classes for ordered pairs, corresponding to one conic lying inside the other (only near the double point for IIN*). For all pairs given here,* $[Q_S = 0]$ *lies inside* $[Q_T = 0]$.

Class	Q_S	Q_T
IaN	$2x^2 - y^2 + z^2$	$3x^2 - 2y^2 + z^2$
IIN	$y^2 - z^2 + xy$	$y^2 - z^2 + 2xy$
IIaN	$y^2 + z^2 + xy$	$y^2 + z^2 + 2xy$
IIIN	$z^2 + x^2 - y^2$	$z^2 + 2x^2 - 2y^2$
IIIaN	$x^2 + y^2 - z^2$	$x^2 + y^2 - 2z^2$
VN	$xz - y^2 - x^2$	$xz - y^2 + x^2$

4. Algebraic invariant theory: an overview. The classical theory of invariants is the study of the intrinsic properties of polynomials and polynomial systems, i.e. properties that are unaffected by a change of variables and are not attached to a specific coordinate system. The study of invariants is intimately linked to the problem of equivalence (can a polyno-

mial system be transformed into another through an appropriate change of coordinates?) and the related problem of the canonical form (find a transformation that puts a polynomial system into a particularly simple form).

Founded by Sylvester and Cayley in the years 1850, the theory of invariants and covariants was explored by Hilbert, Clebsch, Gordan and others around 1880 and rediscovered in the 1960's by Dieudonné and Dixmier, in particular.

Here, we go through the basic elements of the classical theory of algebraic invariants. The interested reader is referred to old ([14, 15, 29]) or modern ([23, 25]) sources for more information on this fascinating subject.

4.1. Invariants of n-ary forms. The invariant theory of forms is a central chapter of classical invariant theory. We state it here by considering the action of $\mathrm{GL}_n(\mathbb{K})$ on forms in $\mathbb{K}^n$, which descends to an action of $\mathrm{PGL}_n(\mathbb{K})$ on hypersurfaces in $\mathbb{P}^{n-1}(\mathbb{K})$.

Let $\mathbb{K}$ be a field of characteristic zero, which will be set to $\mathbb{C}$ or $\mathbb{R}$ later. Let

$$f(x_1, x_2, \ldots, x_n) = \sum a_{i_1 i_2 \cdots i_n} x_1^{i_1} x_2^{i_2} \cdots x_n^{i_n} \tag{4.1}$$

be a n-ary form (i.e. a homogeneous polynomial in n variables) of degree d on $\mathbb{K}$, where $\mathbf{x} = (x_1, \ldots, x_n)$ are coordinates of $\mathbb{K}^n$ and the coefficients $a_{i_1 i_2 \cdots i_n}$ are in $\mathbb{K}$. The sum in (4.1) is over the $\binom{n+d-1}{d}$ n-tuples of nonnegative integers $(i_1, i_2, \ldots, i_n)$ such that $i_1 + i_2 + \cdots + i_n = d$.

Let $(\mathbb{K}^n)^*$ be the set of all linear forms $\gamma : \mathbb{K}^n \to \mathbb{K}$. $(\mathbb{K}^n)^*$ is the vector space dual to $\mathbb{K}^n$. All n-ary forms of degree d form a $\mathbb{K}$-vector space of dimension $\binom{n+d-1}{d}$. It can be identified with $S_d(\mathbb{K}^n)^*$, the d-th symmetric power of $(\mathbb{K}^n)^*$. It implies that f can be identified to the vector $\mathbf{a} = (\ldots, a_{i_1 i_2 \cdots i_n}, \ldots)$ of its coefficients. Thus $f = f(\mathbf{x}) = f(\mathbf{a}, \mathbf{x})$ and $\mathbf{a}$ represent the same element of $S_d(\mathbb{K}^{n*})$.

Define linear forms u_i on $(\mathbb{K}^n)^*$ such that $u_i(x_j) = \delta_{ij}, \forall i, j$. $\mathbf{u} = (u_1, \ldots, u_n)$ are coordinates dual to $\mathbf{x}$, i.e. tangential coordinates. Let $\mathbb{K}[\mathbf{a}, \mathbf{x}, \mathbf{u}]$ be the ring of polynomials in the coefficients of f, the primal variables and the dual variables. $\mathbb{K}[\mathbf{a}, \mathbf{x}, \mathbf{u}]$ is the ring of polynomials over the vector space $\Gamma = S_d(\mathbb{K}^{n*}) \oplus \mathbb{K}^n \oplus (\mathbb{K}^n)^*$. The action of the general linear group $\mathrm{GL}_n(\mathbb{K})$ on $\mathbb{K}^n$ induces a natural action on Γ. For each transformation P of $\mathrm{GL}_n(\mathbb{K})$, represented by a non-singular $n \times n$ matrix with coefficients in $\mathbb{K}$, the action $P : (\mathbf{a}, \mathbf{x}, \mathbf{u}) \mapsto (\overline{\mathbf{a}}, \overline{\mathbf{x}}, \overline{\mathbf{u}})$ is defined by the equations

$$\mathbf{x} = P\,\overline{\mathbf{x}}, \quad f(\mathbf{a}, \mathbf{x}) = f(\overline{\mathbf{a}}, \overline{\mathbf{x}}),$$

the dual coordinates $\mathbf{u}$ undergoing the dual transformation

$$\mathbf{u} = \left(P^{-1}\right)^T \overline{\mathbf{u}}.$$

A polynomial $C \in \mathbb{K}[\mathbf{a}, \mathbf{x}, \mathbf{u}]$ is a *concomitant* of f under the action of $\mathrm{GL}_n(\mathbb{K})$ if

$$C(\overline{\mathbf{a}}, \overline{\mathbf{x}}, \overline{\mathbf{u}}) = (\det P)^w \, C(\mathbf{a}, \mathbf{x}, \mathbf{u}),$$

where the integer w is called the *weight* (or *index*) of C. A concomitant is said to be *(multi-) homogeneous* if it is homogeneous as a polynomial in the coefficients of $\mathbf{a}$, as a polynomial in the variables $x_1, \ldots, x_n$ and as a polynomial in the variables $u_1, \ldots, u_n$. In that case, the total degree of C in the elements of the vector of coefficients $\mathbf{a}$ is called the *degree* of the concomitant. Its total degree in the variables $\mathbf{x}$ is called the *order* of C. Finally, its total degree in the dual variables $\mathbf{u}$ is called the *class* of C.

In many situations, it is desirable to consider not just one but a collection of n-ary forms (not necessarily of the same degrees) $f_i(\mathbf{a}_i, \mathbf{x}), i = 1, \ldots, p$. As before, we consider the natural action of the group $\mathrm{GL}_n(\mathbb{K})$ on the polynomial ring in $\mathbf{x}$, $\mathbf{u}$ and the $\mathbf{a}_i$'s. A polynomial $J \in \mathbb{K}[\mathbf{a}_1, \ldots, \mathbf{a}_p, \mathbf{x}, \mathbf{u}]$ is a *joint concomitant* of the forms $f_1, f_2, \ldots, f_p$ if

$$J(\overline{\mathbf{a}_1}, \ldots, \overline{\mathbf{a}_p}, \overline{\mathbf{x}}, \overline{\mathbf{u}}) = (\det P)^w \, J(\mathbf{a}_1, \ldots, \mathbf{a}_p, \mathbf{x}, \mathbf{u}),$$

for all $P \in \mathrm{GL}_n(\mathbb{K})$. When J is multi-homogeneous, its dependence on the different vectors of coefficients $\mathbf{a}_1, \ldots, \mathbf{a}_p$ is measured by aggregating its degree α_i in each $\mathbf{a}_i$ in a *multi-degree* $(\alpha_1, \ldots, \alpha_p)$. When the forms all have the same degree, any concomitant $C(\mathbf{a}, \mathbf{x}, \mathbf{u})$ induces trivial joint concomitants $J_i(\mathbf{a}_1, \ldots, \mathbf{a}_p, \mathbf{x}, \mathbf{u}) = C(\mathbf{a}_i, \mathbf{x}, \mathbf{u})$ for any $i = 1, \ldots, p$. Vice-versa, if $J(\mathbf{a}_1, \ldots, \mathbf{a}_p, \mathbf{x}, \mathbf{u})$ is a joint concomitant, then its *trace* $C(\mathbf{a}, \mathbf{x}, \mathbf{u}) = J(\mathbf{a}, \ldots, \mathbf{a}, \mathbf{x}, \mathbf{u})$ gives a concomitant of individual forms.

19th-century terminology, due to Sylvester, further classified (joint) concomitants as:

- *invariants*, when the order and the class are 0,
- *covariants*, when the class is 0, but not the order,
- *contravariants*, when the order is 0, but not the class,
- *mixed concomitants*, when neither the order nor the class are 0.

REMARK 4.1. The invariants as we just defined them are *relative* since they are only preserved up to some power of the determinant of the transformation. What is particularly interesting about them is therefore not their value, but rather their vanishing or non-vanishing, as well as possibly their sign (when the weight is even). Note though that given two (relative) invariants of same weight one can construct an *absolute invariant* by dividing them.

Let us now give some well-known examples of concomitants and joint concomitants:

- The most simple covariant of a form f is f itself.
- A trivial example of a mixed concomitant is $\mathcal{U} = \langle \mathbf{x}, \mathbf{u} \rangle = \sum_{i=1}^{n} x_i \, u_i$, which "encodes" the pairing between $\mathbb{K}^n$ and $(\mathbb{K}^n)^*$.

It satisfies the definition of a concomitant for weight 0, independently of any actual form, and is therefore known as the *universal concomitant*.

- The Hessian $\det\left(\frac{\partial^2 f}{\partial x_i \partial x_j}\right)$ of a n-ary form f of degree d is a covariant of f of weight 2, degree n and order $n(d-2)$ [25, Prop. 4.4.2].
- The Jacobian $\frac{\partial(f_1,\ldots,f_n)}{\partial(x_1,\ldots,x_n)} = \det\left(\frac{\partial f_j}{\partial x_i}\right)$ of n n-ary forms f_i of degree d_i, $i = 1,\ldots,n$, is a joint covariant of weight 1, order $\sum d_i - n$ and multi-degree $(1,1,\ldots,1)$ [25, Ex. 4.4.4].
- The resultant of n n-ary forms f_i of degree d_i, which vanishes if and only if the system $\{f_1(\mathbf{x}) = \cdots = f_n(\mathbf{x}) = 0\}$ has a non-zero solution, is a joint invariant of multi-degree $(\delta_1,\ldots,\delta_n)$, where $\delta_i = \prod_{j\neq i} d_j$ [7].
- The discriminant of a form f of degree d in n variables, which vanishes if and only if the projective hypersurface $[f = 0]$ has a singularity, is an invariant of degree $n(d-1)^{n-1}$ [13, Chapter 13]. It is the resultant of the n partial derivatives of f.
- Recall that the dual variety X^* of a subvariety X in $\mathbb{P}^{n-1}$ is the closure in the dual projective space $(\mathbb{P}^{n-1})^*$ of the locus of hyperplanes in $\mathbb{P}^{n-1}$ which are tangent to X at some nonsingular point of X. The operation of taking the dual of a hypersurface defines a contravariant [9, Exercise 5.8].

4.2. Key results. The key results in invariant theory are consequences of the general abstract algebraic results proved by David Hilbert around 1890 [17, 18], which subsequently formed the foundation of modern commutative algebra.

Let $\mathbb{K} = \mathbb{C}$. The *algebra of concomitants* of f is the $\mathbb{C}$-algebra generated by the concomitants of f. The *algebra of invariants* (resp. *covariants*, *contravariants*) of f is the sub-algebra of the algebra of concomitants generated by the invariants (resp. covariants, contravariants) of f. The first fundamental result is the finiteness of bases of these algebras.

THEOREM 4.1 (Hilbert's Finiteness Theorem). *The algebra of concomitants of one or several forms under the action of the general linear group over $\mathbb{C}$ has a finite number of generators. Members of a minimal system of generators of the algebra of concomitants (resp. invariants, covariants, contravariants) are called* fundamental concomitants *(resp.* fundamental invariants*,* fundamental covariants*,* fundamental contravariants*).*

The fundamental concomitants are, in general, not independent (that is, the algebra they generate is not a free algebra). There can exist non-trivial algebraic relations among them, called *syzygies*. The second fundamental result of algebraic invariant theory is another consequence of Hilbert's work:

THEOREM 4.2. *The algebra of concomitants of one or several forms under the action of the general linear group over $\mathbb{C}$ has a finite basis of syzygies.*

4.3. Computing concomitants. Classically, the fundamental problem in the theory of invariants is to calculate the invariants/concomitants and to completely describe them. The calculation itself builds upon two ingredients:

- the observation that a concomitant of a concomitant of a form (resp. a system of forms) is a concomitant of the form (resp. the system of forms) itself (cf. [11, §45] for a proof);
- various formal processes that were developed to this end. We will here describe two such processes: transvection and polarization.

4.3.1. Transvection and Cayley's Ω-process. The process of *transvection* (from the German *Überschiebung*) was described in 1885 by the German mathematician Paul Gordan (cf. [23, Chapter 5] for a detailed presentation and a historical perspective). The basic idea is that concomitants can be recovered by taking the trace of certain naturally defined joint concomitants.

Binary forms. We first focus on binary forms. Transvection is based on an invariant differential operator originally introduced by Cayley. Let us use two independent sets of variables $\mathbf{x}^{(1)} = (x_1^{(1)}, x_2^{(1)})$, $\mathbf{x}^{(2)} = (x_1^{(2)}, x_2^{(2)}) \in \mathbb{C}^2$. We consider the joint action of $\mathrm{GL}_2(\mathbb{C})$ on the Cartesian product space $\mathbb{C}^2 \times \mathbb{C}^2$, given by simultaneous linear transformations $(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) \mapsto (P^{-1}\mathbf{x}^{(1)}, P^{-1}\mathbf{x}^{(2)})$. The second order differential operator

$$\Omega_{\mathbf{x}} = \begin{vmatrix} \frac{\partial}{\partial x_1^{(1)}} & \frac{\partial}{\partial x_1^{(2)}} \\ \frac{\partial}{\partial x_2^{(1)}} & \frac{\partial}{\partial x_2^{(2)}} \end{vmatrix} = \frac{\partial^2}{\partial x_1^{(1)}\,\partial x_2^{(2)}} - \frac{\partial^2}{\partial x_1^{(2)}\,\partial x_2^{(1)}}$$

is known as the Cayley Ω-*process* with respect to the variables $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$. Under the simultaneous linear transformation above, the Ω-process undergoes the transformation

$$\Omega_{\mathbf{x}} \mapsto (\det P)\,\Omega_{\mathbf{x}},$$

proving at once its covariance.

Definition 4.1. *The* k-th transvectant *of two binary forms* $f_1(\mathbf{x})$ *and* $f_2(\mathbf{x})$, $\mathbf{x} = (x_1, x_2)$, *is the function*

$$(f_1, f_2)_k = (\Omega_{\mathbf{x}})^k \big[f_1(\mathbf{x}^{(1)}) f_2(\mathbf{x}^{(2)})\big]\Big|_{\substack{x_1 = x_1^{(1)} = x_1^{(2)} \\ x_2 = x_2^{(1)} = x_2^{(2)}}}.$$

The invariantal nature of transvectants follows from the covariance of the Ω-process.

Theorem 4.3 ([23, Theorem 5.4]). *If* f_1 *and* f_2 *have respective degrees* q_1 *and* q_2 *and respective weights* w_1 *and* w_2, *then* $(f_1, f_2)_k$ *is a joint covariant of* f_1, f_2 *of weight* $w_1 + w_2 + k$ *and order* $q_1 + q_2 - 2k$. *In particular, if* f *has degree* q, $(f, f)_k$ *is a covariant of* f *of weight* k, *degree 2 and order* $2q - 2k$.

Expanding $(\Omega_{\mathbf{x}})^k$ leads to the simple explicit formula

$$(f_1, f_2)_k = \sum_{i=0}^{k} (-1)^i \binom{k}{i} \frac{\partial^k f_1}{\partial x^i \, \partial y^{k-i}} \frac{\partial^k f_2}{\partial y^i \, \partial x^{k-i}}.$$

Observe that $(f_1, f_2)_k = (-1)^k (f_2, f_1)_k$ so the k-th transvectant is symmetric in f_1, f_2 when k is even, skew-symmetric when k is odd.

The process of transvection generalizes certain well-known operations. The first transvectant of f_1 and f_2 is their Jacobian determinant. The second transvectant of a form with itself is a scalar multiple of its Hessian determinant. If $f_1 = a_1 x^2 + a_2 xy + a_3 y^2$, then $-\frac{1}{2}(f_1, f_1)_2 = a_2^2 - 4a_1 a_3$ is the discriminant of f_1, i.e. the simplest example of an algebraic invariant. If in addition $f_2 = a_4 x + a_5 y$, then

$$\frac{1}{2}((f_1, f_2)_1, f_2)_1 = a_1 a_5^2 - a_2 a_4 a_5 + a_3 a_4^2,$$

which is nothing but the resultant of f_1 and f_2.

Starting from one or several forms, concomitants can therefore be produced by taking successive transvectants and composing those transvectants. Actually, a consequence of the First Fundamental Theorem of Invariant Theory is that, over $\mathbb{C}$, all polynomial covariants and invariants[2] of a system of binary forms can be expressed that way [23, Theorems 6.14 and 6.23]. More precisely, in the case of a single binary form f, every concomitant is a linear combination of composed transvectants of the form

$$(\ldots, ((f, f)_{r_1}), f)_{r_2}, \ldots, f)_{r_l}.$$

The determination of an explicit polynomial basis for the covariants of a general binary form is an extremely difficult problem. However, a result due to Stroh and Hilbert constructs an explicit *rational* basis for a form of arbitrary degree (see Theorem 4.4 below). Assume the binary form f has degree $n \geqslant 3$. Define integers l, m such that $l = m - 1$ when $n = 2m$ is even and $l = m$ when $n = 2m+1$ is odd. The only covariant of order 1 of f (up to a constant factor) is f itself. There exist m independent quadratic (i.e., degree 2) covariants

$$S_k = (f, f)_{2k}, \quad k = 1, \ldots, m.$$

S_k has order $2n - 4k$ and weight $2k$. In particular, S_1 is the Hessian of f and, when n is even, $n = 2m$, S_m has order 0, and is therefore an invariant of f. Beyond, there are many cubic (i.e., degree 3) covariants. The most important are:

$$T_k = (f, S_k)_1 = (f, (f, f)_{2k})_1, \quad k = 1, \ldots, l.$$

[2] In the case of binary forms, contravariants are obtained from covariants by a simple determinantal reweighting [23, Example 4.5]. This identification does not carry over to the case of n-ary forms, $n > 2$, for which there exist non-trivial contravariants.

T_k has order $3n-4k-2$ and weight $2k+1$. Then, the Stroh-Hilbert Theorem states that the n fundamental covariants $f, S_1, \ldots, S_m, T_1, \ldots, T_l$ form a rational basis for all the covariants.

THEOREM 4.4 ([23, Theorem 6.32]). *Let C be any polynomial covariant of the binary form f of degree n. Then, for some power $N \geqslant 0$, the covariant $f^N C$ can be written as a polynomial in the n fundamental covariants $f, S_1, \ldots, S_m, T_1, \ldots, T_l$.*

n-ary forms. Transvection can be generalized to n-ary forms (cf. [23, Chapter 10]). Let us give a presentation of this generalization in the case of ternary forms, using the notations of Olver. Let f_1, f_2, f_3 be three forms in the ternary variable $\mathbf{x} = (x_1, x_2, x_3)$. Their tensor product $f_1 \otimes f_2 \otimes f_3$ is identified with the polynomial $f_1(\mathbf{x}^{(1)}) f_2(\mathbf{x}^{(2)}) f_3(\mathbf{x}^{(3)})$ in the three independent ternary variables $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}$ and $\mathbf{x}^{(3)}$. The multiplication map $f_1 \otimes f_2 \otimes f_3 \to f_1 f_2 f_3$ is denoted using the "trace" notation, i.e.

$$\operatorname{tr}\Big(f_1(\mathbf{x}^{(1)}) f_2(\mathbf{x}^{(2)}) f_3(\mathbf{x}^{(3)})\Big) = f_1(\mathbf{x}) f_2(\mathbf{x}) f_3(\mathbf{x}).$$

The Cayley operator $\Omega_{\mathbf{x}}$ acts on such a tensor product by the differential operator

$$\Omega_{\mathbf{x}} = \begin{vmatrix} \frac{\partial}{\partial x_1^{(1)}} & \frac{\partial}{\partial x_1^{(2)}} & \frac{\partial}{\partial x_1^{(3)}} \\ \frac{\partial}{\partial x_2^{(1)}} & \frac{\partial}{\partial x_2^{(2)}} & \frac{\partial}{\partial x_2^{(3)}} \\ \frac{\partial}{\partial x_3^{(1)}} & \frac{\partial}{\partial x_3^{(2)}} & \frac{\partial}{\partial x_3^{(3)}} \end{vmatrix}.$$

If F_1, F_2, F_3 are three forms in the variables $\mathbf{x}$ (primal) and $\mathbf{u}$ (dual), one defines their transvectant as

$$(F_1, F_2, F_3)_n^m = \operatorname{tr}\Big(\Omega_{\mathbf{x}}^n \Omega_{\mathbf{u}}^m \prod_{i=1}^{3} F_i(\mathbf{x}^{(i)}; \mathbf{u}^{(i)})\Big).$$

Then $(F_1, F_2, F_3)_n^m$ is a joint concomitant of the forms F_1, F_2, F_3. If F_1, F_2, F_3 are themselves concomitants of a ternary form f, then $(F_1, F_2, F_3)_n^m$ is a concomitant of the form f, etc. Note that many concomitants and joint concomitants we have already encountered can be embodied as transvectants. For instance, the Jacobian of three ternary forms f_1, f_2, f_2 is (up to a scalar) the transvectant $(f_1, f_2, f_3)_1^0$ and the Hessian of the ternary form f is (up to a scalar) the transvectant $(f, f, f)_2^0$.

4.3.2. Polarization. Knowing the concomitants of a form of $S_d(\mathbb{K}^{n*})$ does not yield the joint concomitants of a collection of forms of $S_d(\mathbb{K}^{n*})$ in an obvious way. Polarization leads to such a construction [33].

Let again $\mathbb{K}[\mathbf{a}, \mathbf{x}, \mathbf{u}]$ be the ring of polynomials over the vector space $\Gamma = S_d(\mathbb{K}^{n*}) \oplus \mathbb{K}^n \oplus (\mathbb{K}^n)^*$. Let $\delta \in \mathbb{K}[\mathbf{a}, \mathbf{x}, \mathbf{u}]$ be multi-homogeneous of total degree d in the coefficients of $\mathbf{a}$. Write

$$\delta(\lambda f_1 + \mu f_2) = \sum_{i=0}^{d} \lambda^{d-i} \mu^i \delta_i(f_1, f_2), \quad \lambda, \mu \in \mathbb{K}, \quad f_1, f_2 \in S_d(\mathbb{K}^{n*}).$$

The δ_i's are called *(partial) polarizations* of δ. We then have the following result (adapted from [19, Sec. 4.5]).

THEOREM 4.5. *If G is any subgroup of $\mathrm{GL}_n(\mathbb{K})$, then δ is a concomitant of forms of $S_d(\mathbb{K}^{n*})$ under the action of G if and only if all δ_i are joint concomitants of pairs of forms of $S_d(\mathbb{K}^{n*})$ under the action of G. Moreover, if δ has order p, class q and weight w, then δ_i has order p, class q, weight w and bidegree $(d-i,i)$.*

Though we stated it here for pairs of forms, the polarization construction can easily be extended to any collection of forms.

4.4. Pencils of conics. A pencil of conics of $\mathbb{P}^2(\mathbb{R})$ is determined by a pencil of real ternary quadratic forms, i.e. a plane through the origin in $S_2(\mathbb{R}^{3^*})$. Let f and g be a pair of real ternary quadratic forms generating the pencil $\lambda f + \mu g, \mathbf{w} = (\lambda, \mu) \in \mathbb{R}^2$, and let $\mathcal{F}$ be the space of all such pairs. Denote by $\mathbb{P}(\mathcal{F})$ the space of pairs of conics. Pencils of conics are generated by elements of $\mathbb{P}(\mathcal{F})$.

The group $\mathrm{GL}_3(\mathbb{R})$ acts naturally on $\mathbb{R}^3$ (and $\mathbb{P}^2(\mathbb{R})$), and thus on $\mathcal{F}$ (and $\mathbb{P}(\mathcal{F})$). The action on $\mathcal{F}$ is by linear substitutions in $\mathbf{x} = (x, y, z)$ (the "space variables") and therefore by simultaneous transformation of f and g. The concomitants with respect to this action are those C such that

$$C(f \circ \theta, g \circ \theta; x, y, z; \lambda, \mu) = (\det \theta)^{w_1} C(f, g; \theta(x, y, z); \lambda, \mu), \quad \forall \theta \in \mathrm{GL}_3(\mathbb{R}).$$

There is also a natural action of the group $\mathrm{GL}_2(\mathbb{R})$ on $\mathbb{R}^2$ (and $\mathbb{P}^1(\mathbb{R})$), and thus on $\mathcal{F}$ (and $\mathbb{P}(\mathcal{F})$). The action on $\mathcal{F}$ is by linear substitutions in f and g and by linear reparameterization of $\mathbf{w}$ (the "pencil variables"). The concomitants with respect to this action are those C such that

$$C(\theta(f, g); x, y, z; \theta(\lambda, \mu)) = (\det \theta)^{w_2} C(f, g; x, y, z; \lambda, \mu), \quad \forall \theta \in \mathrm{GL}_2(\mathbb{R}).$$

We thus have an action of $\mathrm{GL}_3(\mathbb{R}) \times \mathrm{GL}_2(\mathbb{R})$ on $\mathcal{F}$ (and $\mathbb{P}(\mathcal{F})$). Sylvester (1853) calls the concomitants C with respect to this combined action *combinants*. In what follows, we will use the term combinant when C does not depend on the variable $\mathbf{w} = (\lambda, \mu)$, otherwise we will talk about a *combinantal form*.

The action of $\mathrm{GL}_3(\mathbb{R}) \times \mathrm{GL}_2(\mathbb{R})$ on pairs of real ternary quadratic forms descends to an action of $\mathrm{PGL}_3(\mathbb{R})$ on pencils of conics. That this action induces a partitioning of the space of pencils exactly according to the type of their base points follows from the following result.

THEOREM 4.6 ([21]). *Two non-degenerate pencils of conics are equivalent under the action of $\mathrm{PGL}_3(\mathbb{R})$ if and only if they have the same numbers of real and imaginary base points of each multiplicity.*

The roadmap for the rest of this paper is as follows. We start in Section 5 by identifying the concomitants of f and g under the sole action of $\mathrm{GL}_3(\mathbb{C})$ on the space variables. We then move on to the study of combinants of the pencil in Section 6, by first looking at the invariant theory of the

characteristic form. We finally show that the $\mathrm{GL}_3(\mathbb{R})$-invariants of a certain quadratic combinant allow to discriminate in many cases between the orbits of pencils of conics under the action of $\mathrm{PGL}_3(\mathbb{R})$ and we complete the characterization in Section 7.

5. Invariant theory of pairs of ternary quadratic forms. Let Q_S, Q_T be two ternary quadratic forms and $S = (s_i), T = (t_j)$ the associated 3×3 symmetric matrices. We consider from now on the s_i's and t_j's as indeterminates and we work in this section over $\mathbb{C}$. We review the invariant theory of the pair Q_S, Q_T under the action of $\mathrm{GL}_3(\mathbb{C})$, which has been known for over a century.

Note that many results of this section could have been stated for a pair of n-ary quadratic forms. We have chosen to privilege concreteness over generality.

One of the indirect goals of this paper is to minimize the degree of the invariantal quantities discriminating between pencil orbits. In this section and the next, the dependence to input data (the coefficients of the conics) will be measured by a bidegree (d_S, d_T), where d_S (resp. d_T) is the total degree in the coefficients of Q_S (resp. Q_T).

5.1. Invariants. Let us consider the characteristic form of the pencil generated by Q_S, Q_T

$$\mathcal{D}(\lambda, \mu) = \det(\lambda S + \mu T) = a\lambda^3 + b\lambda^2\mu + c\lambda\mu^2 + d\mu^3.$$

Clearly, the coefficients of $\mathcal{D}$ are preserved by a simultaneous congruence transformation $\overline{S} = P^T S P$, $\overline{T} = P^T T P$ of S and T (up to some power of the determinant of the transformation), by property of the determinant. This invariance also follows by polarization, applying Theorem 4.5 with $f_1 = Q_S, f_2 = Q_T$ and taking the discriminant as invariant map δ. a has bidegree $(3,0)$, b has bidegree $(2,1)$, c has bidegree $(1,2)$ and d has bidegree $(0,3)$. All four coefficients have weight 2.

It turns out a, b, c, d are the fundamental joint invariants of Q_S, Q_T.

THEOREM 5.1 ([15, Chapter XIII]). *The coefficients of $\mathcal{D}$ are the generators of the algebra of joint invariants of a pair Q_S, Q_T of ternary quadratic forms under the action of* $\mathrm{GL}_3(\mathbb{C})$. *Every other joint invariant writes as a homogeneous and isobaric polynomial in those coefficients.*

5.2. Non-constant concomitants. Let us now identify the fundamental covariants and contravariants.

5.2.1. Preliminary result. In what follows we will use the following lemma, corollary of Jacobi's fundamental lemma on the minors of an adjoint.

LEMMA 5.1. *Let A be a $n \times n$ matrix. The adjoint $A' = \mathrm{adj}(A)$ of A is such that* $\mathrm{adj}(A') = (\det A)^{n-2} A$.

Before moving on, we prove a preliminary lemma:

LEMMA 5.2. *Let Q_A be a ternary quadratic form, with associated matrix A. Let $A' = \operatorname{adj}(A)$. $Q_{A'}$ is a contravariant of Q_A of degree 2 and weight 2. $Q_{\operatorname{adj}(A')}$ is a covariant of Q_A of degree 4 and weight 2.*

Proof. We consider the effect of a scaling $\mathbf{x} \mapsto \overline{\mathbf{x}} = P^{-1}\mathbf{x}$, with $P = \lambda I$, where I is the identity matrix of size 3. The dual variables undergo the transformation $\overline{\mathbf{u}} = \lambda \mathbf{u}$. In view of $\overline{Q_A} = \overline{\mathbf{x}}^T \overline{A} \overline{\mathbf{x}} = Q_A$, the transformed matrix $\overline{A}$ is $\lambda^2 A$. Therefore $\overline{\operatorname{adj}(A)} = \operatorname{adj}(\overline{A}) = \lambda^4 \operatorname{adj}(A)$ and

$$\overline{Q_{\operatorname{adj}(A)}} = \overline{\mathbf{u}}^T \overline{\operatorname{adj}(A)} \overline{\mathbf{u}} = \lambda^6 Q_{\operatorname{adj}(A)} = (\det P)^2 \, Q_{\operatorname{adj}(A)},$$

showing that $Q_{\operatorname{adj}(A)}$ is a contravariant of weight 2. Its degree follows immediately from the definition of the adjoint.

Since $\operatorname{adj}(A') = (\det A) A$ (Lemma 5.1) and $\det A$ is an invariant of weight 2, $Q_{\operatorname{adj}(A')}$ is a covariant of weight 2. By definition, its degree is 4. □

5.2.2. Covariants. Let $S' = \operatorname{adj}(S)$ and $T' = \operatorname{adj}(T)$ be the adjoint matrices of S and T respectively. Let U be the symmetric matrix defined by

$$\operatorname{adj}(\lambda S' + \mu T') = aS\lambda^2 + U\lambda\mu + dT\mu^2.$$

Let us denote by $Q_U = \mathbf{x}^T U \mathbf{x}$ the associated quadratic form, $\mathbf{x} = (x_0, x_1, x_2)$. Let finally G be the Jacobian

$$G(\mathbf{x}) = \frac{1}{8} \frac{\partial(Q_S, Q_T, Q_U)}{\partial(x_0, x_1, x_2)}.$$

LEMMA 5.3. *Q_U and G are joint covariants of Q_S and Q_T under the action of $\mathrm{GL}_3(\mathbb{C})$. Q_U has bidegree $(2,2)$ and weight 2. G has bidegree $(3,3)$ and weight 3.*

Proof. By polarization, Q_U is a joint concomitant of $Q_{S'}, Q_{T'}$ of bidegree $(1,1)$ in the coefficients of $Q_{S'}, Q_{T'}$ (apply Theorem 4.5 with $f_1 = Q_{S'}$, $f_2 = Q_{T'}$ and the duality map as concomitant map δ). Since $Q_{S'}$ and $Q_{T'}$ are themselves joint contravariants of Q_S, Q_T of bidegree $(2,0)$ and $(0,2)$ respectively, we conclude by composition that Q_U is joint covariant of Q_S, Q_T of bidegree $(2,2)$. In view of Lemma 5.2, its weight is 2.

The covariance of the Jacobian has already been mentioned in Section 4.1. Since Q_U is a joint covariant of Q_S and Q_T, G is a also a joint covariant by composition. Since Q_S, Q_T, Q_U have bidegree $(1,0), (0,1)$ and $(2,2)$ respectively, this adds up to G having bidegree $(3,3)$. Finally, note that

$$\frac{\partial(\,\cdot\,,\,\cdot\,,\,\cdot\,)}{\partial(\overline{x_0}, \overline{x_1}, \overline{x_2})} = (\det P) \, \frac{\partial(\,\cdot\,,\,\cdot\,,\,\cdot\,)}{\partial(x_0, x_1, x_2)}$$

when $\mathbf{x} = (x_0, x_1, x_2)$ undergoes the transformation $\mathbf{x} = P\overline{\mathbf{x}}$. Since the weights of Q_S, Q_T, Q_U are 0, 0 and 2 respectively, the weight of G is 3. □

It turns out that the joint invariants and covariants we identified above form a complete, irreducible system.

THEOREM 5.2 ([15, Chapter XIII]). *Every joint covariant of Q_S, Q_T can be expressed as an integral function of $a, b, c, d, Q_S, Q_T, Q_U$ and G.*

REMARK 5.1. The joint covariants Q_S, Q_T, Q_U and G are not algebraically independent. They satisfy a fundamental syzygy which, when $Q_S = Q_T = 0$, reduces to

$$G^2 = Q_U^3.$$

5.2.3. Contravariants. Let now E be the symmetric matrix defined by

$$\operatorname{adj}(\lambda S + \mu T) = \lambda^2 S' + \lambda\mu E + \mu^2 T'.$$

As previously, let us denote the quadratic form associated to E by $Q_E = \mathbf{u}^T E \mathbf{u}$, $\mathbf{u} = (u_0, u_1, u_2)$. Let also H be the Jacobian

$$H(\mathbf{u}) = \frac{1}{8} \frac{\partial(Q_{S'}, Q_{T'}, Q_E)}{\partial(u_0, u_1, u_2)}.$$

Proceeding as for Lemma 5.3, we prove the following.

LEMMA 5.4. *Q_E and H are joint contravariants of Q_S and Q_T under the action of $\mathrm{GL}_3(\mathbb{C})$. Q_E has bidegree $(1,1)$ and weight 2. H has bidegree $(3,3)$ and weight 5.*

Proof. By polarization Q_E is a joint contravariant of Q_S, Q_T of bidegree $(1,1)$ (apply Theorem 4.5 with $f_1 = Q_S, f_2 = Q_T$ and the duality map as contravariant map δ). Also, in view of Lemma 5.2, Q_E has weight 2.

Since Q_E is a joint contravariant of $Q_{S'}, Q_{T'}$, H is a joint contravariant by composition. Since $Q_{S'}, Q_{T'}, Q_E$ have bidegree $(2,0), (0,2)$ and $(1,1)$ respectively, this adds up to H having bidegree $(3,3)$. Finally,

$$\frac{\partial(\cdot, \cdot, \cdot)}{\partial(\overline{u_0}, \overline{u_1}, \overline{u_2})} = (\det P)^{-1} \frac{\partial(\cdot, \cdot, \cdot)}{\partial(u_0, u_1, u_2)},$$

proving at once the weight of H is 5 (sum of the weights of $Q_{S'}, Q_{T'}, Q_E$ minus 1). ∎

As before, the system of joint contravariants formed by $a, b, c, d, Q_{S'}, Q_{T'}$, Q_E and H is both complete and irreducible.

THEOREM 5.3 ([15, Chapter XIII]). *Every joint contravariant of Q_S, Q_T can be expressed as an integral function of $a, b, c, d, Q_{S'}, Q_{T'}, Q_E$ and H.*

We don't discuss here the case of mixed concomitants (cf. [15, Chapter XIII]).

5.3. Linking covariants and contravariants. There are multiple relationships between the matrix forms of covariants and contravariants. Two of particular importance in this paper are given in the following lemma, the proof of which is deferred to Appendix A.

LEMMA 5.5. *The following identities hold:*

$$\begin{aligned}\operatorname{adj}(\lambda S' + \mu T' + \kappa E) &= aS\lambda^2 + dT\mu^2 \\ &\quad + (cS + bT - U)\kappa^2 + U\lambda\mu + (bS + aT)\lambda\kappa + (dS + cT)\mu\kappa,\end{aligned}$$

$$\begin{aligned}\operatorname{adj}(\lambda S + \mu T + \kappa U) &= S'\lambda^2 + T'\mu^2 \\ &\quad + (bdS' + acT' - adE)\kappa^2 + E\lambda\mu + (cS' + aT')\lambda\kappa + (dS' + bT')\mu\kappa.\end{aligned}$$

6. Combinants of pairs of ternary quadratic forms. Let us now focus on the combined action of $\mathrm{GL}_3(\mathbb{C})$ on the space variables and $\mathrm{GL}_2(\mathbb{C})$ on the pencil variable. In other words, we begin our investigation of combinants of the pencil.

6.1. Invariant combinants. We start with the case of combinants that do not depend on the space variables, i.e. the invariants of the pencil. Fortunately, they are easy to identify:

THEOREM 6.1 ([1, Theorem 7], [24, Theorem 14]). *The ring of invariants of pencils of quadratic hypersurfaces of* $\mathbb{P}^{n-1}(\mathbb{C})$ *under the action of* $\mathrm{SL}_n(\mathbb{C})$ *is isomorphic to the ring of invariants of binary forms of degree* n *under the action of* $\mathrm{SL}_2(\mathbb{C})$.

More precisely, the invariants of a pencil of quadratic hypersurfaces are the invariants of its characteristic form. Going back to the case of a pencil of conics, the characteristic form $\mathcal{D}$ is a cubic:

$$\mathcal{D}(\lambda, \mu) = \det(\lambda S + \mu T) = a\lambda^3 + b\lambda^2\mu + c\lambda\mu^2 + d\mu^3.$$

So in accordance with Theorem 6.1 let us review the well-known invariant theory of cubics under the action of $\mathrm{GL}_2(\mathbb{C})$ (see for instance [6]).

The non-constant covariants of $\mathcal{D}$ are $\mathcal{D}$ itself, its Hessian (degree 2, order 2)

$$\mathcal{H}(\lambda, \mu) = -\frac{1}{8}(\mathcal{D}, \mathcal{D})_2 = (b^2 - 3ac)\lambda^2 + (bc - 9ad)\lambda\mu + (c^2 - 3bd)\mu^2,$$

and the Jacobian of $\mathcal{D}$ and $\mathcal{H}$ (degree 3, order 3)

$$\mathcal{G}(\lambda, \mu) = (\mathcal{D}, \mathcal{H})_1 = (2b^3 + 27a^2d - 9abc)\lambda^3 + \cdots.$$

The only invariant of $\mathcal{D}$ under the action of $\mathrm{GL}_2(\mathbb{C})$ is its discriminant, which is also (up to a scalar) the discriminant of $\mathcal{H}$:

$$\Delta = \frac{1}{6}(\mathcal{H}, \mathcal{H})_2 = b^2c^2 - 4ac^3 - 4b^3d - 27a^2d^2 + 18abcd.$$

Δ is equal to the classical formulation of the discriminant of a cubic, that is $-\frac{1}{a}\operatorname{Res}(\mathcal{D},\mathcal{D}')$, if one considers for a moment $\mathcal{D}$ as a polynomial in λ (assuming $a \neq 0$). $\Delta > 0$ when $\mathcal{D}$ has three simple real roots, $\Delta < 0$ when $\mathcal{D}$ has a real root and two complex roots, and $\Delta = 0$ when $\mathcal{D}$ has a multiple root.

We note that these covariants are not algebraically independent since they satisfy the following syzygy:

$$4\mathcal{H}^3 - \mathcal{G}^2 - 3^3\Delta\mathcal{D}^2 = 0.$$

Clearly, the characteristic form $\mathcal{D}$ and its covariants are combinants of the pencil. Since Δ has bidegree $(6,6)$ in Q_S and Q_T (so even degree in each of the input conics), its sign is also invariant on each real orbit of pencils.

6.2. Covariant combinants. Let us now focus on the construction of combinants of the pencil other than the invariants and covariants of the characteristic form $\mathcal{D}$. Note that our goal is to uncover some combinants that we have identified as being relevant to distinguish between pencil orbits, not to make an exhaustive study (see Todd [27] for an attempt at determining the complete irreducible system of combinants).

We will here start with simple combinantal forms and compute some of their (joint) $\mathrm{GL}_2(\mathbb{C})$-invariants to generate combinants of the pencil. We have already encountered several combinantal forms: the characteristic form $\mathcal{D}$, its Hessian $\mathcal{H}$ and

$$\mathcal{R}(\lambda,\mu) = \lambda Q_S + \mu Q_T.$$

The quadratic form dual to $\mathcal{R}$ is also a combinantal form of the pencil. In view of Section 5, this dual form is:

$$\mathcal{S}(\lambda,\mu) = \lambda^2 Q_{S'} + \lambda\mu Q_E + \mu^2 Q_{T'}.$$

The discriminant of $\mathcal{S}$ (as a binary form) is a (contravariant) combinant of the pencil:

$$\mathcal{B} = -\frac{1}{2}(\mathcal{S},\mathcal{S})_2 = Q_E^2 - 4Q_{S'}Q_{T'}.$$

$\mathcal{B}$ has bidegree $(2,2)$ in the coefficients of the input conics.

We now determine the main combinant that will be used in the rest of the paper:

PROPOSITION 6.1. *The quadratic covariant*

$$\mathcal{J} = -cQ_S - bQ_T + 3Q_U$$

is a combinant of bidegree $(2,2)$ *of the pencil of conics generated by* Q_S *and* Q_T*. Its discriminant is equal to* Δ*.*

Proof. Let us first take the second transvectant of $\mathcal{H}$ and $\mathcal{S}$:

$$\frac{1}{2}(\mathcal{H},\mathcal{S})_2 = 2(c^2-3bd)Q_{S'} + (9ad-bc)Q_E + 2(b^2-3ac)Q_{T'}.$$

This quadratic contravariant is a combinant of the pencil. Let us compute the adjoint of the associated matrix K, using Lemma 5.5. This gives:

$$\operatorname{adj}(K) = \Delta(-cS - bT + 3U).$$

Let J be the matrix $-cS - bT + 3U$. By duality, $Q_{\Delta J} = \Delta\mathcal{J}$ is a covariant combinant of the pencil. Δ being itself an invariant of the pencil (Section 6.1), we conclude that $\mathcal{J}$ is a combinant of the pencil (Q_S, Q_T).

Using the second identity of Lemma 5.5, we show that

$$\operatorname{adj}(J) = -K.$$

Taking adjoint on both sides, we get (Lemma 5.1):

$$\begin{aligned}\operatorname{adj}(\operatorname{adj}(J)) &= (\det J)\, J,\\ &= \operatorname{adj}(K) = \Delta J,\end{aligned}$$

which proves, by Weyl's Principle (see Appendix A), that $\det J = \Delta$. □

Now we know exactly what the invariants of a real ternary quadratic form under the action of $\mathrm{GL}_3(\mathbb{R})$ are and since J has even degree in S and T, we conclude:

COROLLARY 6.1. *The inertia of $\mathcal{J}$ is an invariant of the pencil orbits under the action of* $\mathrm{PGL}_3(\mathbb{R})$.

The inertia $\operatorname{in}(\mathcal{J})$ of $\mathcal{J}$ is found by inspection of the characteristic polynomial of the associated matrix:

$$\det(\ell I - J) = \ell^3 - \operatorname{tr}(\mathcal{J})\ell^2 + \gamma(\mathcal{J})\ell - \Delta.$$

J being symmetric, Descartes' Rule of Signs [2] tells us that the number of positive eigenvalues of J is the number of sign changes in the sequence

$$(+, -\operatorname{sgn}(\operatorname{tr}(\mathcal{J})), \operatorname{sgn}(\gamma(\mathcal{J})), -\operatorname{sgn}(\Delta))$$

and the number of negative eigenvalues is the number of sign changes in the sequence

$$(-, -\operatorname{sgn}(\operatorname{tr}(\mathcal{J})), -\operatorname{sgn}(\gamma(\mathcal{J})), -\operatorname{sgn}(\Delta)).$$

Note that $\operatorname{tr}(\mathcal{J})$ has bidegree $(2,2)$ and $\gamma(\mathcal{J})$ has bidegree $(4,4)$. They are however not invariant quantities of the pencil.

6.3. Geometric meaning. Let us now give a geometric interpretation of the vanishing or non-vanishing of some of the combinants we have identified. Though this interpretation is in no way needed for separating the orbits (we can simply treat combinants as algebraic quantities and evaluate them on the representatives of each orbit), linking simple combinants with geometric loci associated with the pencil clearly illuminates the problem.

The interested reader is referred to [26, 27] for a deeper understanding of the beautiful geometry behind pencils of conics.

PROPOSITION 6.2. *Δ vanishes on all orbits except Orbits I, Ia and Ib.*

Proof. The discriminant Δ does not vanish exactly when $\mathcal{D}$ has only simple roots. It is known to be equivalent to the intersection of the conics being generic, i.e. made of four simple points over the complex numbers [5]. □

PROPOSITION 6.3 (Hesse). *$\mathcal{H}$ vanishes identically iff $\mathcal{D}$ has a triple root or $\mathcal{D} \equiv 0$.*

COROLLARY 6.2. *$\mathcal{H}$ vanishes identically on Orbits V, VI, VIa, VII and VIII+.*

Assume for a moment that the intersection is made of four (real or complex) points, i.e. the characteristic form $\mathcal{D}$ has three simple roots ℓ_i. Each of the singular form $\ell_i Q_S + Q_T$ has rank 2. The three associated dual forms have therefore rank 1, i.e. they are the squares of three linear forms in dual space. Associated to these three forms, by duality, are three points p_i of $\mathbb{P}^2(\mathbb{C})$ forming the vertices of a triangle. It can be shown that $G = 0$ is the equation of the sides of this triangle (and $H = 0$ is the product of the three lines dual to the p_i's). Since each side $p_i p_j$ of the triangle is the polar of opposite vertex p_k with respect to every conic of the pencil, G is often called the *self-polar triangle covariant*.

G and H are invariantly attached to the pencil of conics, not just to ternary quadratic forms generating it. That they are combinants of the pencil can easily be seen algebraically: when Q_S (say) is replaced by a linear combination of Q_S and Q_T, Q_U is replaced by a combination of Q_S, Q_T and Q_U (since they are the only quadratic covariants of pairs of ternary quadratic forms) and therefore the Jacobian G is unchanged. The same goes for H.

PROPOSITION 6.4. *G and H vanish identically if the conic pencil generated by Q_S and Q_T contains a conic of rank less than or equal to 1.*

Proof. Let ℓ_0 be the pencil parameter of a conic of rank 1 or less. The adjoint of the corresponding matrix vanishes identically, i.e.

$$\operatorname{adj}(\ell_0 S + T) = \ell_0^2 S' + \ell_0 E + T' \equiv 0. \tag{6.1}$$

The three contravariants $Q_{S'}, Q_E, Q_{T'}$ are thus not linearly independent and their Jacobian, i.e. H, vanishes identically. Now multiply the left-hand side of (6.1) to the left by aS and to the right by dT:

$$aS\,\mathrm{adj}\,(\ell_0 S+T)\,dT = \mathrm{adj}\,(S')\,\mathrm{adj}\,(\ell_0 S+T)\,\mathrm{adj}\,(T'),$$
$$= \mathrm{adj}\,(T'(\ell_0 S+T)S') = \mathrm{adj}\,(dS'+a\ell_0 T') \equiv 0.$$

Expanding the last term gives

$$ad(dS+\ell_0 U+a\ell_0^2 T)\equiv 0.$$

The three covariants Q_S, Q_U, Q_T are thus not linearly independent and their Jacobian, i.e. G, vanishes identically. □

COROLLARY 6.3. *G vanishes identically on Orbits III, IIIa, V, VI, VIa, VII and VIII+.*

PROPOSITION 6.5. *$\mathcal{B}$ vanishes identically when the intersection $Q_S \cap Q_T$ is 1-dimensional.*

Proof. The line $L : \langle \mathbf{u}, \mathbf{x}\rangle = 0$ is tangent to $\lambda Q_S + \mu Q_T$ if $\mathbf{u}$ is such that

$$\mathbf{u}^T\,\mathrm{adj}\,(\lambda S+\mu T)\mathbf{u} = \mathcal{S}(\lambda,\mu) = 0.$$

L meets the pencil of conics in a pencil of binary quadratic forms and the section of the conic $\lambda Q_S + \mu Q_T$ by the line is singular if the line is tangent to the conic. The combinantal form $\mathcal{S}$ is therefore a constant multiple of the characteristic form of the pencil of binary quadratic forms. The discriminant of this characteristic form, i.e. $\mathcal{B}$, is zero exactly when the quadratic forms of the pencil all share a common root, or in other words when the conics of the pencil $\lambda Q_S + \mu Q_T$ have a 1-dimensional intersection. □

COROLLARY 6.4. *$\mathcal{B}$ vanishes identically on Orbits VI, VIa and VII.*

REMARK 6.1. It would be interesting to have a better grasp of the relationship between the vanishing or non-vanishing of simple combinants and the characteristics of pencil orbits – multiplicities of roots of $\mathcal{D}$, ranks of attached conics – that naturally stem from Segre's classification of pairs of quadratic forms over the complex numbers (see [5]).

7. Distinguishing between pencil orbits. Let us now see how the invariants and combinants we have identified help distinguish between pencil orbits.

Let us start with the case $\mathcal{D}(\lambda,\mu) \not\equiv 0$. In light of Theorem 6.1, our first "discriminant" will be the only invariant of the pencil under the action of $\mathrm{PGL}_3(\mathbb{C})$, i.e. Δ. Computing the sign of Δ on the representative of each pencil orbit given in Table 1, we immediately obtain the following:

- $\Delta > 0$: I, Ia;
- $\Delta < 0$: Ib;
- $\Delta = 0$: II, IIa, III, IIIa, IV, V.

There is thus nothing left to do when $\Delta < 0$.

When $\Delta > 0$, let us compute the inertia of $\mathcal{J}$:

- I: $\mathrm{in}\,(\mathcal{J}) = (3,0)$,
- Ia: $\mathrm{in}\,(\mathcal{J}) = (1,2)$.

To distinguish between Orbits I and Ia, we thus introduce the following condition:

$$Z_{\mathcal{J}} := \operatorname{tr}(\mathcal{J}) > 0 \wedge \gamma(\mathcal{J}) > 0.$$

$Z_{\mathcal{J}}$ is true on Orbit I and false on Orbit Ia.

Let us now move to the case $\Delta = 0$. We evaluate the inertia of $\mathcal{J}$ on each of the remaining orbits:

- II: $\operatorname{in}(\mathcal{J}) = (2, 0)$,
- IIa: $\operatorname{in}(\mathcal{J}) = (1, 1)$,
- III: $\operatorname{in}(\mathcal{J}) = (1, 0)$,
- IIIa: $\operatorname{in}(\mathcal{J}) = (0, 1)$,
- IV: $\operatorname{in}(\mathcal{J}) = (1, 0)$,
- V, VI, VIa: $\operatorname{in}(\mathcal{J}) = (0, 0)$.

We thus see that the sign of $\gamma(\mathcal{J})$ gives another discriminant:

- $\gamma(\mathcal{J}) > 0$: II;
- $\gamma(\mathcal{J}) < 0$: IIa;
- $\gamma(\mathcal{J}) = 0$: III, IIIa, IV, V, VI, VIa.

When $\gamma(\mathcal{J}) = 0$, the sign of the trace of $\mathcal{J}$ provides a further boundary line:

- $\operatorname{tr}(\mathcal{J}) > 0$: III, IV;
- $\operatorname{tr}(\mathcal{J}) < 0$: IIIa;
- $\operatorname{tr}(\mathcal{J}) = 0$: V, VI, VIa.

To distinguish between the remaining cases, it is enough to note that $\mathcal{H} \not\equiv 0$ and $G \equiv 0$ on Orbit III, while $\mathcal{H} \equiv 0$ and $G \not\equiv 0$ on Orbit IV. Then, $\mathcal{B} \not\equiv 0$ on Orbit V and $\mathcal{B} \equiv 0$ on Orbits VI and VIa. Finally, $\mathcal{D} \leqslant 0$ on Orbit VI and $\mathcal{D} \geqslant 0$ on Orbit VIa.

Let us go briefly over the cases where $\mathcal{D}(\lambda, \mu) \equiv 0$. Here again the inertia of $\mathcal{J}$ allows to distinguish:

- VII: $\operatorname{in}(\mathcal{J}) = (1, 0)$;
- VIII+: $\operatorname{in}(\mathcal{J}) = (0, 0)$.

Consequently, $\operatorname{tr}(\mathcal{J}) \neq 0$ on Orbit VII and $\operatorname{tr}(\mathcal{J}) = 0$ otherwise.

We therefore obtain the decision tree displayed in Figure 2. It should be noted that the predicate of largest degree we need to evaluate to determine the pencil orbit is Δ and therefore determining the type of the intersection of two real projective conics can be achieved with predicates of bidegree at most $(6, 6)$ in the input conics. This improves upon the results of Briand [4] who gives predicates of largest bidegree $(13, 6)$ for distinguishing between orbits.

8. Isotopy classes and inside-outside test. Deciding in which rigid isotopy class a pair of conics lies knowing the orbit of the pencil they generate is simple. Recall that a pair of conics is in class N if the conics are on the same arc of the pencil (i.e., you can go from one to the other without encountering a singular conic of the pencil), in class S otherwise. For two different classes N and S to exist inside an orbit, it is thus necessary that the characteristic form $\mathcal{D}$ has all its roots real (three simple real roots, or a

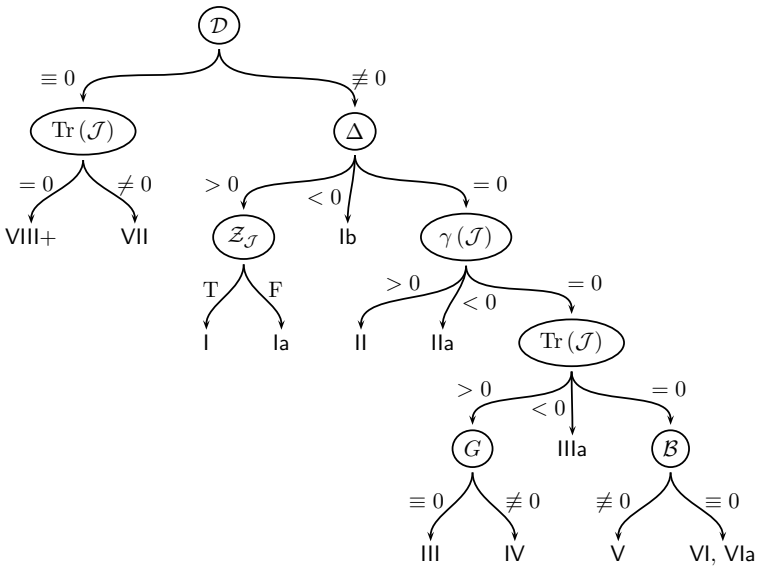


FIG. 2. *Characterization of the* $\mathrm{PGL}_3(\mathbb{R})$ *orbit of a pencil of real projective conics.*

simple root and a double root). Class IaN therefore corresponds to $\mathcal{D}$ having all its roots real, non-zero and of the same sign. As observed by Briand [4], this in turn implies, by Descartes' Rule of Signs, that

$$ac > 0 \wedge bd > 0.$$

Now assume a given pair of conics lies in one of the classes IaN, IIN, IIaN, IIIN, IIIaN and VN. We want to decide which conic is inside the other. We are therefore looking for an invariant characteristic of the ordered pair of conics. In other words, while we previously considered *symmetric* quantities (i.e., unchanged by the substitutions $Q_S \to Q_T, Q_T \to Q_S, a \to d, b \to c, c \to b, d \to a, Q_U \to Q_U, G \to -G$), we now look for *antisymmetric* quantities of invariantal nature. The simplest antisymmetric covariant that is isobaric and has the same degree in both input conics (i.e., 2) is

$$Q_A = cQ_S - bQ_T.$$

Note that it has even degree in the coefficients of both conics, so is really an invariant of the algebraic conics under the action of $\mathrm{GL}_3(\mathbb{C})$, not just of the forms defining them. So its inertia is an invariant of the ordered pair $([Q_S], [Q_T])$ under the action of $\mathrm{GL}_3(\mathbb{R})$. Let us show here that it is also invariant inside each rigid isotopy class.

The proof technique is the same as in [4]. Let

$$Q_S = Q_S^0 + t_1 Q_T^0, \quad g = Q_S^0 + t_2 Q_T^0,$$

where Q_S^0 and Q_T^0 are taken (in that order) from Table 1. For the cases IIN, IIaN, IIIN and IIIaN, the inner conic is the one "closer" to $[Q_S^0]$, i.e. the one whose parameter t_i has smaller absolute value. Consider the determinant of the matrix A associated to Q_A:

$$\det A = ac^3 - db^3.$$

We prove the following:

LEMMA 8.1. *Let* $([Q_S], [Q_T])$ *be an ordered pair of proper non-empty conics. If* $([Q_S], [Q_T])$ *is in Class* IIN, IIaN *or* IIIN *then* $[Q_S = 0]$ *is inside* $[Q_T = 0]$ *iff* $\det A > 0$. *If* $([Q_S], [Q_T])$ *is in Class* IIIaN *then* $[Q_S = 0]$ *is inside* $[Q_T = 0]$ *iff* $\det A < 0$.

To prove this, we evaluate $\det A$ on (Q_S, Q_T). For instance for IIa, we have

$$\det A = \frac{1}{256} t_1^2 t_2^2 (t_2 - t_1)^2 (t_2^2 - t_1^2),$$

so $\det A > 0$ when $|t_2| > |t_1|$, proving the result. Note that Lemma 8.1 corrects several sign errors in [4].

The case of Class Ia is harder (Briand [4] fails to realize that the sign of $\det A$ also gives the answer in this case and resorts to other methods). For the representatives of Class Ia in Table 1, we have

$$\det A = (t_1 t_2 - 1)(t_2 - t_1)^3 (t_1 + t_2 + t_1 t_2)(t_1 + t_2 + 1)$$

whose sign is not immediately clear. Note that Q_S and Q_T write as

$$(1 + t_i)x^2 + y^2 + t_i z^2,$$

so they are non-empty and proper iff either $t_i \in (-1, 0)$ or $t_i \in (-\infty, -1)$. For the two conics to be on the same arc (and therefore in the same isotopy class) their parameters must both be in the same interval. $[Q_S^0]$ and $[Q_T^0]$ both being of inertia $(2, 0)$ (a point), separated at $t_i = -1$ by a conic of inertia $(1, 1)$ (a pair of lines), in each case the inside conic is the one "closer" to the extremity of the arc corresponding to a point.

Assume the t_i's are in the first interval, i.e. $t_i \in (-1, 0)$. Then $t_1 t_2 < -t_1 < 1$ and

$$t_1 + t_2 + t_1 t_2 < t_2 < 0.$$

Note in addition that

$$(t_2 - t_1)(t_1 + t_2 + 1) = t_2^2 - t_1^2 + t_2 - t_1. \tag{8.1}$$

Here the inside conic is the one "closer" to $[Q_S]$, i.e. the one whose parameter t_i has smaller absolute value. If $|t_2| > |t_1|$, which means $t_1 > t_2$, (8.1) is negative so putting things together yields $\det A < 0$.

Assume now the t_i's are in the second interval, i.e. $t_i \in (-\infty, -1)$. Then $t_1 t_2 > 1$ and

$$t_1 + t_2 + 1 < 0.$$

We rewrite $\det A$ as follows:

$$\det A = (t_1 t_2)^4 (t_1 t_2 - 1) \left(\frac{1}{t_1} - \frac{1}{t_2}\right)^3 \left(\frac{1}{t_1} + \frac{1}{t_2} + 1\right) (t_1 + t_2 + 1).$$

Note that

$$\left(\frac{1}{t_1} - \frac{1}{t_2}\right) \left(\frac{1}{t_1} + \frac{1}{t_2} + 1\right) = \left(\frac{1}{t_1}\right)^2 - \left(\frac{1}{t_2}\right)^2 + \frac{1}{t_1} - \frac{1}{t_2}. \tag{8.2}$$

Here the inside conic is the one "closer" to $[Q_T]$, i.e. the one with smallest $|1/t_i|$. If $|1/t_2| > |1/t_1|$, which means $1/t_1 > 1/t_2$, (8.2) is positive so putting things together yields $\det A < 0$.

We conclude as follows:

LEMMA 8.2. *Let* $([Q_S], [Q_T])$ *be an ordered pair of proper non-empty conics in Class* IaN. *Then* $[Q_S = 0]$ *is inside* $[Q_T = 0]$ *iff* $\det A < 0$.

The last case, that of Class VN, is treated by Briand [4]. He shows that $\det A$ vanishes identically on Class VN but considers the trace $\operatorname{tr}(A)$ of A and proves that $[Q_S = 0]$ is inside $[Q_T = 0]$ iff $\operatorname{tr}(A) < 0$.

Overall, we have proved that the inside-outside test for pairs of conics in classes IaN, IIN, IIaN, IIIN, IIIaN and VN can be carried out with predicates of bidegree at most $(6, 6)$ in the input conics (thereby finishing the proof of Theorem 1.1). This improves upon the results of Briand [4] who gives predicates of largest bidegree $(17, 8)$ for the same task.

REMARK 8.1. There must be a fundamental reason for the invariance of Q_A here which our "pedestrian" proof fails to grasp.

REMARK 8.2. Consider the rigid isotopy classes CN for pairs of conics that split into two classes CN_i for ordered pairs. Briand [3] shows that rigid isotopy is preserved by duality (i.e. by the mapping sending a conic to its dual). More precisely, he shows that the classes for ordered pairs are exchanged: $\mathsf{CN}_1 \leftrightarrow \mathsf{CN}_2$. In other words, if $[Q_S = 0]$ is inside $[Q_T = 0]$, then $[Q_{T'} = 0]$ is inside $[Q_{S'} = 0]$.

This is coherent with the invariance of $\det A$. Indeed, by Lemma A.2, the coefficients of the characteristic form of the dual pencil $([Q_{S'}], [Q_{T'}])$ are $a' = a^2, b' = ac, c' = bd, d' = d^2$. Therefore, for the dual pair of conics, we have:

$$\det A' = a' c'^3 - d' b'^3 = a^2 d^2 (b^3 d - c^3 a) = -a^2 d^2 \det A$$

and therefore the sign is reversed, as expected.

9. Examples. Let us now apply the decision procedure we just outlined to two examples.

9.1. Example 1. Consider the pencil generated by the two conics of equation:

$$\begin{cases} Q_S : x^2 - 12xy - 12xz - 3y^2 + 10yz + 12z^2 = 0, \\ Q_T : -3x^2 - 20xy - 8xz - 7y^2 + 14yz + 11z^2 = 0. \end{cases}$$

We wish to determine in which orbit this pencil lies.

The characteristic form of the pencil is:

$$\mathcal{D}(\lambda, \mu) = -25\lambda^3 - 100\lambda^2\mu - 125\lambda\mu^2 - 50\mu^3.$$

The discriminant of this form is $\Delta = 0$.

Let us now construct the combinant $\mathcal{J}$. We start by determining the covariant Q_U. For this, we compute the adjoints of S and T, then we extract the coefficient of $\lambda\mu$ in the equation:

$$\operatorname{adj}(\lambda S' + \mu T') = aS\lambda^2 + U\lambda\mu + dT\mu^2.$$

This gives, in equation form:

$$Q_U = 25x^2 + 1100xy + 800xz + 325y^2 - 850yz - 875z^2.$$

Now we can form the combinant $\mathcal{J}$:

$$\mathcal{J} = -cQ_S - bQ_T + 3Q_U = -25(2x + 2y - z)^2.$$

As expected, the discriminant of this ternary form (which is Δ) is zero. In addition, we see that $\mathcal{J}$ has rank 1, i.e. $\gamma(\mathcal{J}) = 0$. Finally, $\operatorname{tr}(\mathcal{J}) < 0$. The decision tree of Fig. 2 tells us that the pencil lies in Orbit IIIa. As a consequence, the intersection of the two conics is empty of real point.

Since $ac > 0 \wedge bd > 0$, where as before a, b, c, d denote the coefficients of the characteristic form, the pair of conics is in rigid isotopy class IIIaN in view of the results of Section 8. In addition

$$\det A = ac^3 - db^3 = -1171875 < 0,$$

implying by the results of Section 8 that $[Q_S = 0]$ is inside $[Q_T = 0]$.

9.2. Example 2. We now consider the pencil generated by the following pair of forms:

$$\begin{cases} Q_S : x^2 + 2txy - 4txz + ty^2 + z^2 = 0, \\ Q_T : x^2 + 2xy - 2yz - 3z^2 = 0. \end{cases}$$

Q_S is a function of a parameter t (time, say) and we wish to know for which values of this parameter the two conics have empty intersection. In other words, we want to know when the pencil generated by Q_S and Q_T falls in Orbit Ia or IIIa.

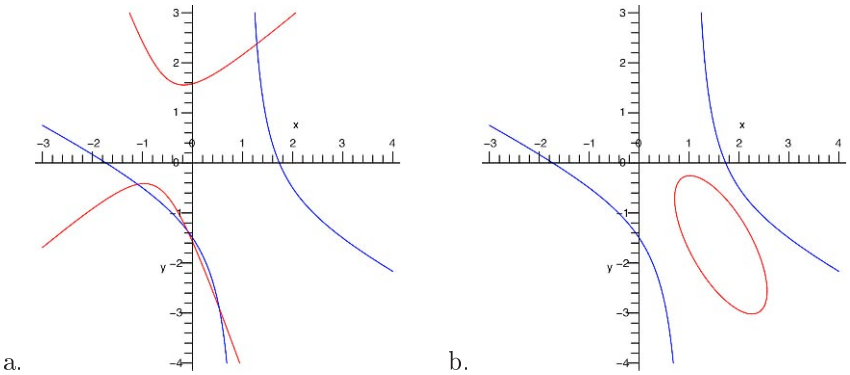


FIG. 3. *Intersection of a deforming conic (in red) with a fixed conic (in blue). a. Case where the intersection is made of four real points. b. Case where the intersection is empty of real point.*

The fundamental invariants of the pair of forms are:

$$a = t - t^2 - 4t^3, \quad b = -4t + 7t^2, \quad c = -2 + 7t, \quad d = 2.$$

We then obtain the discriminant Δ of the characteristic form, a polynomial of degree 6 in t:

$$\Delta = t(32 - 124t - 360t^2 + 1536t^3 + 676t^4 - 3639t^5).$$

$\Delta > 0$ for $t \in (-0.582, -0.333)$ and $t \in (0, 0.504)$. On each of these intervals, the morphology of the intersection does not change (otherwise, we would by continuity go through a singular intersection, corresponding to a zero of Δ). The condition

$$Z_{\mathcal{J}} = \mathrm{tr}\,(\mathcal{J}) > 0 \wedge \gamma\,(\mathcal{J}) > 0$$

is thus either globally verified on each interval, or globally violated. It is therefore sufficient to verify it for an arbitrary value of t inside each interval. We first determine $\mathcal{J}$ in matrix form

$$\begin{pmatrix} -1 - 12t + 26t^2 & -6t + 7t^2 & 3 - 7t + 2t^2 \\ -6t + 7t^2 & -4t + 8t^2 & -4t + 19t^2 \\ 3 - 7t + 2t^2 & -4t + 19t^2 & -1 - 10t + 33t^2 \end{pmatrix}.$$

From this we compute

$$\gamma\,(\mathcal{J}) = -8 + 72t + 20t^2 - 804t^3 + 916t^4, \quad \mathrm{tr}\,(\mathcal{J}) = -2 - 26t + 67t^2.$$

For the first interval, we evaluate for instance at $t = t_0 = -1/2$:

$$\gamma\,(\mathcal{J})_{t_0} = \frac{475}{4}, \quad \mathrm{tr}\,(\mathcal{J})_{t_0} = \frac{111}{4}.$$

The condition $Z_{\mathcal{J}}$ is thus satisfied on this interval, which means the intersection is made of four simple real points (Fig. 3.a). For the second interval, let us evaluate at $t = t_1 = 1/2$:

$$\gamma(\mathcal{J})_{t_1} = -\frac{41}{4}, \quad \operatorname{tr}(\mathcal{J})_{t_1} = \frac{7}{4}.$$

This time, the condition $Z_{\mathcal{J}}$ is violated, and the intersection is empty on the whole interval (cf. Fig. 3.b).

The only other way for the intersection to be empty is for the pencil to belong to Orbit IIIa. But this implies in particular that Δ and $\gamma(\mathcal{J})$ vanish simultaneously. A little calculation shows that the resultant of these two polynomials is not zero, and therefore this case can not happen.

Now assume t is inside the second interval we have identified above, i.e. $t \in (0, 0.504)$. b is negative on the whole interval and d is positive on the whole interval, so $bd < 0$ implying by the results of Section 8 that the ordered pair of conics is in isotopy class IaS.

10. Conclusion. In this paper we have proved, using tools from algebraic invariant theory, that the projective type of the intersection of a pair of conics, their rigid isotopy class and which of the two is inside the other (when applicable) can be determined with predicates of bidegree at most $(6, 6)$ in their coefficients. The results largely improve upon previous approaches.

It would be interesting to prove or disprove that bidegree $(6, 6)$ is optimal. It is very likely that it indeed is since $(6, 6)$ is the bidegree of the discriminant (and only invariant) of the characteristic form whose vanishing separates generic from non-generic cases. There may however be room for improvement in the description of individual orbits, i.e. there may exist, for some orbits, a set of predicates characterizing the orbits having lower degree than the one we have identified.

In applications, it may be interesting to restrict the range of input conics to certain types, for instance parabolas only. This may lead to fewer orbits, less predicates to evaluate and possibly predicates of lower degree. Etayo *et al.* [12] have for instance characterized the relative position of two ellipses using tools from real algebraic geometry (Sturm-Habicht sequences). Future work will be devoted to understanding how their results relate to ours and if/how the characterization of the positions of type-specific conics can be obtained from our invariant-based approach.

Another obvious direction of research is the extension of the results of this paper to quadrics in $\mathbb{P}^3(\mathbb{R})$, which involves studying the invariant theory of pairs of quaternary quadratic forms. This may however prove quite difficult since the algebra of combinants is largely more complex than in the case of conics.

Acknowledgement. The author wishes to acknowledge Emmanuel Briand and the referees for their help in improving this manuscript.

APPENDIX

A. Proof of Lemma 5.5. We here give a proof of Lemma 5.5, split across several lemmas.

To get rid of certain artificial restrictions appearing on coefficients, we will frequently call upon the following result, due to Hermann Weyl [33], which we call Weyl's Principle:

THEOREM A.1 (Principle of irrelevance of algebraic inequalities). *A polynomial $F(x, y, \ldots)$ over a field $\mathbb{K}$ vanishes identically if it vanishes numerically for all sets of rational values $x = \alpha, y = \beta, \cdots$ subject to a number of algebraic inequalities*

$$R_1(\alpha, \beta, \cdots) \neq 0, \quad R_2(\alpha, \beta, \cdots) \neq 0, \quad \ldots$$

Recall that the coefficients of the characteristic form of the pencil generated by Q_S, Q_T are a, b, c, d, that the quadratic covariants of the pencil are Q_S, Q_T, Q_U and that the quadratic contravariants of the pencil are $Q_{S'}, Q_{T'}, Q_E$.

LEMMA A.1. *We have the following identities:*

$$aE = bS' - S'TS', \quad dE = cT' - T'ST'.$$

Proof. Let us form the matrix product of $\lambda S + \mu T$ with its adjoint:

$$(\lambda S + \mu T)\ \mathrm{adj}\,(\lambda S + \mu T) = \det(\lambda S + \mu T)\, I,$$

where I is the identity matrix of size 3. Developing this product, we obtain the following identities by term-by-term identification of coefficients:

$$SE + TS' = b\, I, \quad ST' + TE = c\, I.$$

It follows that:

$$S'TS' = S'(b\, I - SE) = bS' - aE.$$

Similarly, $T'ST' = cT' - dE$. □

LEMMA A.2. *The characteristic form of the dual pencil is:*

$$\det(\lambda S' + \mu T') = a^2\lambda^3 + ac\lambda^2\mu + bd\lambda\mu^2 + d^2\mu^2.$$

Proof. We write:

$$S(\lambda S' + \mu T')T = d\mu S + a\lambda T.$$

Taking determinant on both sides leads to:

$$a\, d\, \det(\lambda S' + \mu T') = a(d\mu)^3 + b(d\mu)^2(a\lambda) + c(d\mu)(a\lambda)^2 + d(a\lambda)^3.$$

We obtain the desired result by first assuming that $ad \neq 0$ and then applying Weyl's Principle. □

LEMMA A.3. *We have the following identities:*

$$U = cS - ST'S = bT - TS'T.$$

Proof. We now form the matrix product of $\lambda S' + \mu T'$ with its adjoint:

$$(\lambda S' + \mu T')\ \mathrm{adj}\,(\lambda S' + \mu T') = \det(\lambda S' + \mu T')\, I.$$

We now need to develop this product and use Lemma A.2. We obtain the desired identities by term-by-term identification of the coefficients, after multiplying to the left by S as in the proof of Lemma A.1. □

Let us now finish the proof of Lemma 5.5.

Proof. We first note that:

$$\begin{aligned} ad\,(\lambda S' + \kappa E) &= (a\lambda + b\kappa)dS' - d\kappa S'TS', && \text{(by Lemma A.1)}\\ &= (a\lambda + b\kappa)T'TS' - d\kappa S'TS',\\ &= ((a\lambda + b\kappa)T' - d\kappa S')TS'. \end{aligned}$$

Taking adjoint on both sides, using the development of $\mathrm{adj}\,(\lambda S' + \mu T')$ and applying Lemma A.3, we obtain:

$$\begin{aligned} a^2d^2\,\mathrm{adj}\,(\lambda S' + \kappa E) &= \mathrm{adj}\,(S')\,\mathrm{adj}\,(T)\,\mathrm{adj}\,((a\lambda + b\kappa)T' - d\kappa S'),\\ &= aST'(aS(-d\kappa)^2\\ &\qquad + U(-d\kappa)(a\lambda + b\kappa) + dT(a\lambda + b\kappa)^2),\\ &= a^2d^2(aS\lambda^2 + (bS + aT)\lambda\kappa + (cS + bT - U)\kappa^2). \end{aligned}$$

We therefore have obtained the coefficients of $\lambda^2, \lambda\kappa$ and κ^2 in the development of $\mathrm{adj}\,(\lambda S' + \mu T' + \kappa E)$, the restriction on a and d being lifted by application of Weyl's Principle. The coefficient of $\lambda\mu$ being known, we obtain the missing coefficients by computing, by symmetry, $\mathrm{adj}\,(\mu T' + \kappa E)$.

The second identity of Lemma 5.5 is obtained similarly. □

REFERENCES

[1] D. AVRITZER AND R. MIRANDA. Stability of pencils of quadrics in $\mathbb{P}^4$. *The Boletin de la Sociedad Matematica Mexicana*, III Ser. 5(2):281–300, 1999.

[2] S. BASU, R. POLLACK, AND M.-F. ROY. *Algorithms in Real Algebraic Geometry*, Volume 10 of *Algorithms and Computation in Mathematics*. Springer-Verlag, Berlin, 2003.

[3] E. BRIAND. Duality for couples of conics. Unpublished, 2005.

[4] E. BRIAND. Equations, inequations and inequalities characterizing the configurations of two real projective conics. *Applicable Algebra in Engineering, Communication and Computing*, 18(1–2):21–52, 2007.

[5] T. BROMWICH. *Quadratic Forms and Their Classification by Means of Invariant Factors*. Cambridge Tracts in Mathematics and Mathematical Physics, 1906.

[6] J. CREMONA. Classical invariants and 2-descent on elliptic curves. *Journal of Symbolic Computation*, 31(1/2):71–87, 2001.

[7] C. D'Andrea and A. Dickenstein. Explicit formulas for the multivariate resultant. *Journal of Pure and Applied Algebra*, 164:59–86, 2001.

[8] O. Devillers, A. Fronville, B. Mourrain, and M. Teillaud. Algebraic methods and arithmetic filtering for exact predicates on circle arcs. *Comput. Geom. Theory Appl.*, 22:119–142, 2002.

[9] I. Dolgachev. *Lectures on Invariant Theory*. Cambridge University Press, 2003. London Mathematical Society Lecture Note Series, Volume 296.

[10] L. Dupont, D. Lazard, S. Lazard, and S. Petitjean. Near-optimal parameterization of the intersection of quadrics: II. A classification of pencils. *Journal of Symbolic Computation*, 43(3):192–215, 2008.

[11] E. Elliott. *An Introduction to the Algebra of Quantics*. Clarendon Press, Oxford, 1913.

[12] F. Etayo, L. González-Vega, and N. del Rio. A new approach to characterizing the relative position of two ellipses depending on one parameter. *Computer Aided Geometric Design*, 23(4):324–350, 2006.

[13] I. Gelfand, M. Kapranov, and A. Zelevinsky. *Discriminants, Resultants and Multidimensional Determinants*. Birkhäuser, Boston, 1994.

[14] O. Glenn. *A Treatise on the Theory of Invariants*. Ginn and Company, Boston, 1915.

[15] J.H. Grace and A. Young. *The Algebra of Invariants*. Cambridge University Press, 1903.

[16] D.A. Gudkov. Plane real projective quartic curves. In *Topology and Geometry - Rohlin Seminar*, Volume 1346 of *Lecture Notes in Math.*, pages 341–347. Springer-Verlag, 1988.

[17] D. Hilbert. Über die Theorie der algebraischen Formen. *Math. Ann.*, 36:473–534, 1890.

[18] D. Hilbert. Über die vollen Invariantensysteme. *Math. Ann.*, 42:313–373, 1893.

[19] H. Kraft and C. Procesi. Classical Invariant Theory, A Primer, 2000. Lecture Notes.

[20] T. Lam. *The Algebraic Theory of Quadratic Forms*. W.A. Benjamin, Reading, MA, 1973.

[21] H. Levy. *Projective and Related Geometries*. The Macmillan Co., New York, 1964.

[22] Y. Liu and F.-L. Chen. Algebraic conditions for classifying the positional relationships between two conics and their applications. *J. Comput. Sci. Technol.*, 19(5):665–673, 2004.

[23] P.J. Olver. *Classical Invariant Theory*. Cambridge University Press, 1999.

[24] D. Pervouchine. *Orbits and Invariants of Matrix Pencils*. PhD thesis, Moscow State University, 2002.

[25] B. Sturmfels. *Algorithms in Invariant Theory*. Texts and Monographs in Symbolic Computation. Springer-Verlag, 1993.

[26] J. Todd. *Projective and Analytical Geometry*. Pitman, London, 1947.

[27] J.A. Todd. Combinant forms associated with a pencil of conics. *Proc. Lond. Math. Soc.*, II Ser. 50:150–168, 1948.

[28] C. Tu, W. Wang, B. Mourrain, and J. Wang. Using signature sequences to classify intersection curves of two quadrics. Computer Aided Geometric Design, 2008, to appear.

[29] H.W. Turnbull. *The Theory of Determinants, Matrices and Invariants*. Blackie (London, Glasgow), 1929.

[30] F. Uhlig. A canonical form for a pair of real symmetric matrices that generate a nonsingular pencil. *Linear Algebra and Its Applications*, 14:189–209, 1976.

[31] W. Wang and R. Krasauskas. Interference analysis of conics and quadrics. In *Topics in Algebraic Geometry and Geometric Modeling*, Volume 334 of *Contemp. Math.*, pages 25–36. Amer. Math. Soc., 2003.

[32] W. Wang, J. Wang, and M.-S. Kim. An algebraic condition for the separation of two ellipsoids. *Computer Aided Geometric Design*, 18(6):531–539, 2001.
[33] H. Weyl. *The Classical Groups, Their Invariants and Representations*. Princeton University Press, 1946.

A NOTE ON PLANAR HEXAGONAL MESHES

WENPING WANG* AND YANG LIU*

Abstract. We study the geometry and computation of free-form hexagonal meshes with planar faces (to be called *P-Hex meshes*). Several existing methods are reviewed and a new method is proposed for computing P-Hex meshes to approximate a given surface. The outstanding issues with these methods and further research directions are discussed.

Key words. Planar hexagonal meshes, Dupin indicatrix, polyhedral approximation.

1. Introduction. A hexagonal mesh with planar faces is a discrete polyhedral surface in 3D whose faces are planar hexagons and whose vertices have degree 3. It will be abbreviated as the *P-Hex mesh* through out this paper. P-Hex meshes are used in architecture design of glass/steel panel structures and provide a useful representation for various special surfaces, such as minimal surfaces or constant mean curvature surfaces [2], in discrete differential geometry. (See Figure 1.) There are several existing methods for computing a P-Hex mesh to approximate a given shape. We will review these methods to motivate further research. In addition, we will study the geometric properties of P-Hex meshes and present a new method for computing P-Hex meshes. We will consider robust computation of offset surfaces specific to P-Hex meshes.

The requirement on face planarity of P-Hex meshes arises naturally in modeling of glass/steel panel structures in architecture. Each flat glass panel, represented by a hexagonal face, is framed by beams which are joined at nodes represented by the vertices of the mesh. The node complexity, defined as the number of beams joined at a node, is a major consideration in manufacturing cost. Since their vertices have degree 3, the P-Hex meshes offer the simplest node complexity compared with meshes with planar quadrilateral faces or triangle meshes [6].

Only a closed surface of genus 1 (e.g. a torus) may be tiled with a P-Hex mesh. On a closed surface of genus 0, faces other than hexagons must be used, provided that all the vertices are of degree 3 (for example, see Figure 2). Assuming that only hexagons and pentagons are allowed, it is easy to show that there have to be exactly 12 pentagons. A typical soccer ball is an example of such a tiling of a surface of genus 0, which has 12 pentagons and 20 hexagons.

Two concepts important to the study of P-Hex meshes are *conjugate curve network* and the *Dupin indicatrix* of a surface. Consider a point p

*Department of Computer Science, University of Hong Kong, Pokfulam Road, Hong Kong SAR, P.R. China. The work was partially supported by the National Key Basic Research Project of China under 2004CB318000.

I.Z. Emiris et al. (eds.), *Nonlinear Computational Geometry*, The IMA Volumes in Mathematics and its Applications 151, DOI 10.1007/978-1-4419-0999-2_9,

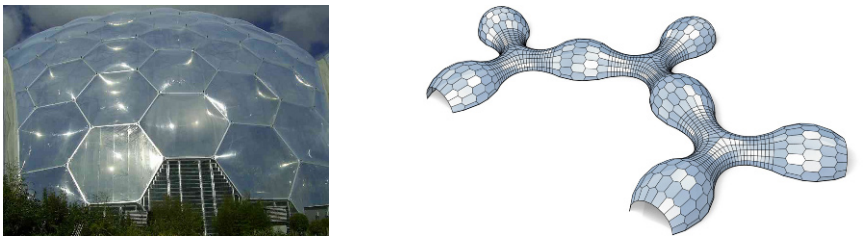


FIG. 1. *Left: a geodome constructed using a P-Hex mesh in the Eden Project in UK; right: the convex parts of this model are constant mean curvature (CMC) surfaces modeled by P-Hex meshes [6].*

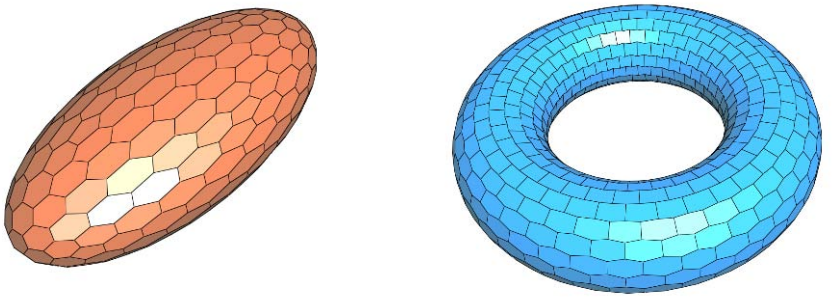


FIG. 2. *Left: an ellipsoid tiled with P-Hex faces and 12 planar pentagons; right: a torus tiled entirely with P-Hex faces.*

on a surface S. Let $T_p(S)$ denote the 2D space of tangent vectors to S at p. Then the differential of the Gauss map, which is the differential $\mathrm{d}N$ of the unit normal vector N of S at p, defines a self-adjoint linear map on $T_p(S)$. Two vectors v and w are *conjugate* at p if the inner product $\langle \mathrm{d}N(v), w\rangle = 0$ [7]; note that this relationship is symmetric, since $\mathrm{d}N$ is a self-adjoint. In particular, at a point p on a developable surface the unique ruling direction at p is conjugate to any other direction. A *conjugate curve network* on S consists of two families of curves, F_1 and F_2, on S such that at any point $p \in S$ there is a unique curve in F_1 and a unique curve in F_2 passing through p and the tangent vector to the curve in F_1 and the tangent vector to the curve in F_2 are conjugate.

Suppose that we have a 2D local coordinate system on $T_p(S)$ with the x and y axes aligned with the principal curvature directions of S at p. Then the *Dupin indicatrix* is a conic defined by $\kappa_1 x^2 + \kappa_2 y^2 = \pm 1$, where κ_1, κ_2 are principal curvatures [7]. Specifically, when p is an elliptic point, assuming that the two principal curvatures $\kappa_1 > 0$ and $\kappa_2 > 0$ by changing the orientation of the surface if necessary, the Dupin indicatrix

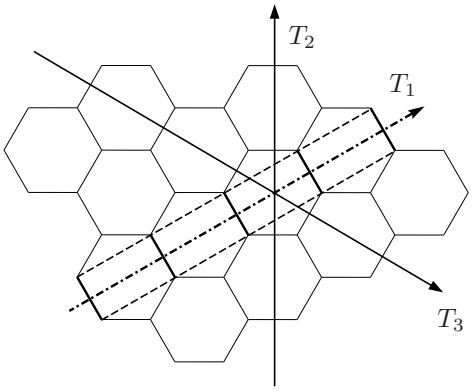


FIG. 3. *Three strips on a P-Hex mesh along the directions T_1, T_2 and T_3.*

is the ellipse $\kappa_1 x^2 + \kappa_2 y^2 = 1$. When p is a hyperbolic point, the Dupin indicatrix consists of two hyperbolas $\kappa_1 x^2 + \kappa_2 y^2 = \pm 1$ having the same pair of asymptotic lines. When p is a parabolic point, assuming that $\kappa_1 \neq 0$ and $\kappa_2 = 0$, the Dupin indicatrix is the pair of lines $\kappa_1 x^2 = \pm 1$. The Dupin indicatrix is not defined at a planar point, where $\kappa_1 = \kappa_2 = 0$. Within the above 2D x-y coordinate system in $T_p(S)$, the Dupin indicatrix has the polar representation $\rho = \pm 1/\sqrt{|\kappa(\theta)|}$, where $\kappa(\theta)$ is the normal curvature of S in the direction of the vector $(\cos\theta, \sin\theta)^T$.

We will take an asymptotic approach in our subsequent analysis. We assume a sequence of P-Hex meshes converging to a surface S, with each hex face h converging to the tangent plane of S at the center of h. An asymptotic analysis is useful to designing numerical methods in practice when a P-Hex mesh is a close approximation to a smooth surface and the faces of the P-Hex meshes are sufficiently small.

A P-Hex mesh comprises three families of developable strips (see Figure 3). Here a developable strip is a surface consisting of a sequence of planar faces joining consecutively along line segments. A developable trip has a central curve formed by the polygon connecting the centers of consecutive hex faces of the strip. Note that the edges between consecutive faces of a developable strip are the discrete rulings of the developable strip. Therefore these edges are conjugate to the direction of the central curve, as a consequence of the discrete analogue of the classic result for smooth developable surfaces [7].

At the center of a hex face h, the central curves of the three strips containing h define three directions. Meanwhile, we assume that in the limit each pair of opposite edges of h are parallel, that is, h has central symmetry. Therefore, the three pairs of opposite edges of h define another three directions at the center of h. It follows that the first three directions are conjugate to the latter three, respectively. Hence, the hex face is con-

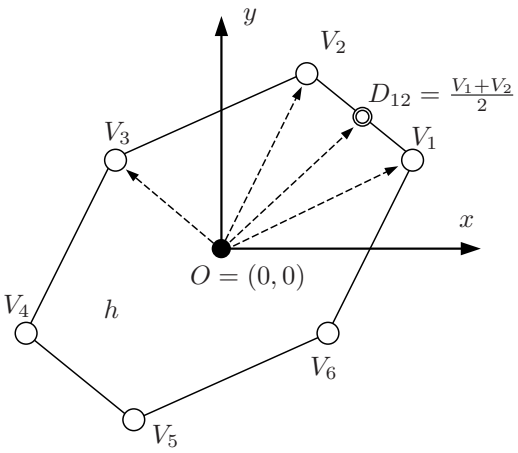


FIG. 4. *Illustration for the proof of Theorem 1.*

strained by a homothetic copy of the Dupin indicatrix, as summarized by the following theorem. (A *homothetic copy* of a shape is the image of the shape under uniform scaling and translation.)

Theorem 1: *Let M be a P-Hex mesh converging to a surface S, with each face of M converging to the tangent plane of S at the center of the face. Let h be a face of M with its center being a point O on the surface S. In the limit, h is inscribed to a homothetic copy of the Dupin indicatrix of the surface S at O.*

Proof. Refer to Figure 4. First we set up a local 2D coordinate system on the tangent plane Γ of the surface $\mathcal{S}$ at O, with the coordinate axes in the principal curvature directions at O. In the limit we can assume that h lies on the tangent plane Γ, with its center at O. Due to its central symmetry, the hex face h is uniquely determined by its vertex vectors V_1, V_2, V_3, with $V_4 = -V_1$, $V_5 = -V_2$, and $V_6 = -V_3$.

In the above 2D coordinate system on the tangent plane Γ, denote $V_i = (\ell_i \cos\theta_i, \ell_i \sin\theta_i)^T$, $i = 1, 2, 3$, subject to that $\ell_i > 0$, $\theta_1 < \theta_2 < \theta_3$ and $\theta_3 - \theta_1 < \pi$. Consider any two consecutive vertices, say V_1 and V_2 (Figure 4). Note that the strip along the central curve direction $D_{12} = (V_1 + V_2)/2$ is conjugate to the ruling direction $V_2 - V_1$ with respect to the inner product $\langle X, Y\rangle \equiv X^T \mathrm{diag}(\kappa_1, \kappa_2) Y$ defined by the second fundamental form. Therefore,

$$\kappa_1(\ell_1^2 \cos^2\theta_1 - \ell_2^2 \cos^2\theta_2) + \kappa_2(\ell_1^2 \sin^2\theta_1 - \ell_2^2 \sin^2\theta_2) = 0.$$

It then follows from Euler's theorem that

$$\kappa(\theta_1)\ell_1^2 - \kappa(\theta_2)\ell_2^2 = 0, \tag{1.1}$$

where $\kappa(\theta_j)$ is the normal curvature in the direction $(\cos\theta_j, \sin\theta_j)^T$, $j = 1, 2$. Comparing (1.1) with the polar representation of the Dupin indicatrix

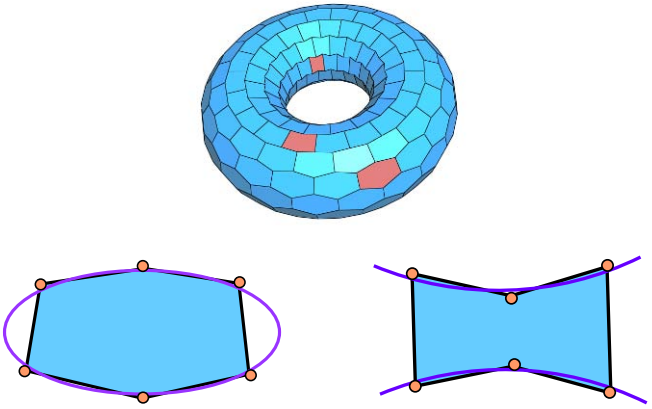


FIG. 5. *Upper: A P-Hex mesh tiling a torus; lower left: a convex P-Hex face is in a first approximation inscribed to a homothetic copy of the Dupin conic, which is an ellipse, when $K > 0$; lower right: a concave P-Hex face is in a first approximation inscribed to a homothetic copy of the Dupin conic, which is a hyperbola, when $K < 0$.*

given previously, we conclude that the six vertices of h lie on a homothetic copy of the Dupin indicatrix. □

The above analysis indicates that convex planar hex faces appear only in an elliptic region of a surface, where the Gaussian curvature $K > 0$ and the Dupin indicatrix is an ellipse, and the P-Hex faces are concave hexagons in a hyperbolic region, where $K < 0$, since they are inscribed to hyperbolas (see Figure 5). Even in an elliptic region, we in general cannot expect to have P-Hex faces to be regular hexagons, since the Dupin indicatrix is in general not a circle.

2. Existing methods. In [3] stereographic projection is used to map a power diagram of a set of points in 2D, which is an extension of Voronoi diagram, onto an ellipsoid to form a polyhedral surface with planar faces. If the faces of the power diagram are hexagons, then a P-Hex mesh approximating the ellipsoid will be generated. This method cannot be extended to other types of quadrics, such as a hyperboloid of one-sheet, or more general free-form surfaces.

An elegant and effective approach to computing a P-Hex mesh is based on projective duality, which establishes a relationship between a triangular mesh and a P-Hex mesh. In fact, this relationship has been used to derive subdivision rules for P-Hex meshes from subdivision rules for triangle meshes [4]. When applying this approach to generating a P-Hex mesh from a triangle mesh, it suffers from the lack of robustness common to several other existing methods.

Recall that projective duality is a transformation that maps a plane $aX + bY + cZ + dW = 0$ in 3D prime space into the point $Q(X, Y, Z, W)^T$

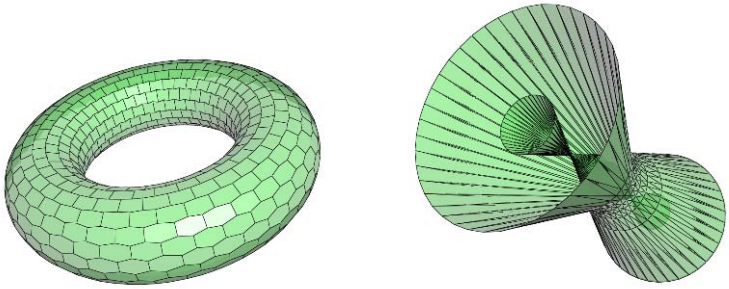


FIG. 6. *Left: a P-Hex mesh approximating a torus; right: the projective dual of the torus with its triangulation corresponding to the P-Hex mesh in the left figure.*

in homogenous coordinates in dual space, where Q is a given symmetric matrix. With an affine specialization, we consider the particular duality that maps a plane not passing through the origin, in the form $ax + by + cz + 1 = 0$ in primal space, to a point $(a, b, c)^T$ in dual space. Under this mapping, a surface S is mapped to another surface S', called the *dual* of S, consisting of points corresponding to the tangent planes of S. Clearly, a P-hex mesh approximating a surface S is dual to a regular triangle mesh approximating S', the dual of S, with each hex face being dual to a degree 6 vertex of the triangle mesh. This property suggests the following method for computing a P-Hex mesh. Given a surface S, first compute the dual S' of S, then compute a regular triangulation of S', and finally map this triangulation to a P-Hex mesh approximating S.

However, there are three major problems with this approach: 1) projective duality may have high metric distortion and parabolic points of S give rise to singular points on S' (see Figure 6). These make it difficult to compute a good triangle mesh on S'; 2) Under projective duality the correspondence between the points of S and the points of S' is often not one-to-one. This makes it difficult to map a triangulation of S' in dual space back to a P-Hex mesh of S in primal space; 3) It is not clear what kind of triangle meshes of S' correspond to P-Hex meshes whose faces are free of self-intersection (see Figure 7). As the consequence of these drawbacks, the method based on projective duality cannot be used to generate P-Hex mesh tiling a free-form surface S. Moreover, even when S is convex the method often generates invalid P-Hex meshes with self-intersecting faces, as illustrated in Figure 7.

The method in [1] uses the supporting function defined over the Gaussian sphere of a free-form surface S to compute a P-Hex mesh. The idea is to first obtain a piecewise linear approximation of the supporting function over a triangulation of the Gaussian sphere. Then it can be shown that the surface determined by this piecewise linear supporting function is a P-Hex mesh approximating the surface S.

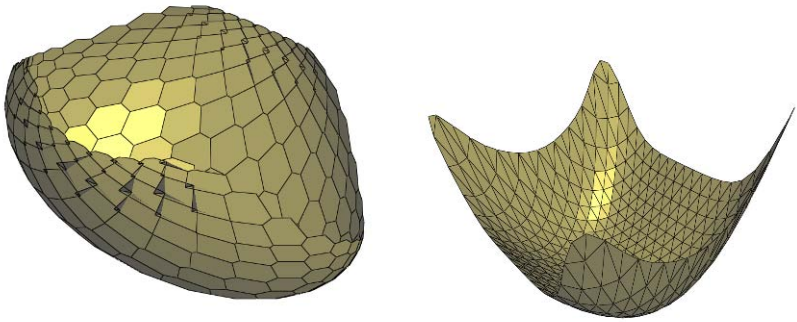


FIG. 7. *Left: a convex surface S approximated by a self-intersecting P-Hex mesh; right: the dual of S represented by a triangular mesh corresponding to the P-Hex mesh in the left figure.*

Consider a tangent plane $ax + by + cz + 1 = 0$ of S. With the support function, this tangent plane is represented by the unit normal vector $(a/m, b/m, c/m) \in S^2$, where $m = (a^2 + b^2 + c^2)^{1/2}$, and its distance from the origin $(0, 0, 0)$ to the plane, which is $1/m$. Therefore, the graph of the support function over S^2 can be represented by the point $\mathbf{p} = \frac{1}{m}(a/m, b/m, c/m) = (a/m^2, b/m^2, c/m^2)$. We recognize that $\mathbf{p}$ is the inversion with respect to the sphere S^2 of the point $(a, b, c)^T$, which is the dual point of the tangent plane $ax + by + cz + 1 = 0$. Hence, the support function can be regarded as the composition of the duality and the spherical inversion with respect to S^2. Because of this, the method in [1] has the same limitations of the other methods based on projective duality. As a consequence, it can only be applied to a surface patch with all elliptic points ($K > 0$) or all hyperbolic points ($K < 0$), and even in these simple cases it often produces invalid P-Hex meshes.

The concept of *parallel meshes* is proposed in [6] for defining and computing various types of offset surfaces of a mesh surface. It may also be used for computing P-Hex mesh for simple surfaces, such as a surface patch with $K > 0$ everywhere or $K < 0$ everywhere. With this method, for example, the convex parts of the model in Figure 1(b) are modeled as P-Hex meshes parallel to a convex Koebe mesh [6]. A restrictive assumption here is that there is already a P-Hex mesh H available, and a new P-Hex mesh H' approximating a given surface S will then be generated as a parallel mesh of H. Moreover, again in this case the P-Hex mesh H' often contains faces with self-intersection.

This review shows that no existing method is capable of computing a valid P-Hex mesh of free-form shape. So it will be a breakthrough if a robust method can be developed for computing valid free-form P-Hex

meshes. There are two major problems we must address to achieve this goal. The first is *generality* — we hope to have a method capable of computing a P-Hex surface approximating any free-form surface, with elliptic, hyperbolic, and parabolic regions all existing on the same surface. The second is *validity* — as the most basic requirement by practical applications, we need to ensure that the faces of the computed P-Hex mesh are free of self-intersection. In addition, from the design point of view, there is a need to explore the full flexibility of P-Hex meshes to allow fine control of the shape and size of the hexagonal faces of a P-Hex mesh.

3. A new method. We will propose a simple method for computing P-Hex meshes. This method has two main steps – we first compute an initial hexagonal mesh that is close to a P-Hex mesh and then use local perturbation to produce the final P-Hex mesh.

As input we start with a conjugate curve network on a target surface S to be approximated (see the left figure of Fig. 8). Sampling these two families of curves, we obtain a quad mesh that is nearly a planar quad mesh [5]. Then we shift every other row of the quad mesh to form a brick-wall layout, which consists of nearly planar hexagonal faces. (See the middle figure of Fig. 8.)

In the second step we use nonlinear optimization to locally perturb the above hexagonal mesh into an exact P-Hex mesh. Note that every 4-point subset of the 6 vertices of a hex face defines a tetrahedron. Obviously, the hex face is planar if and only if the volumes of all these tetrahedra are zero. Therefore, the constraints of our optimization are that the volumes of all the tetrahedra of all the hex faces be zero. To prevent the vertices of the hex mesh from shifting away from the target surface S, we minimize an objective function defined as the sum of the squared distances of the mesh vertices to S. Thus, we end up with a constrained nonlinear least squares problem. We have implemented a penalty method to solve this problem and obtained satisfactory results. The flow of processing is illustrated using a torus in Figure 8. Figure 9 shows the computation of a P-Hex mesh approximating an open surface patch containing different types of curved regions.

The new method can handle a general surface which contains both regions of positive curvature and regions of negative curvature. Empirically, the face self-intersection are removed by using appropriate sampling sizes of the input conjugate curve network. However, a clear theoretical understanding of face self-intersection and a guaranteed practical measure for avoiding it are still missing. We also need to point out that the new method only produces approximately planar hexagonal faces due to its minimization nature, while previously duality-based methods produce exactly planar hexagonal faces.

4. Offset mesh. A closely related issue is the computation of the offset surfaces of P-Hex meshes, which are demanded for modeling multi-

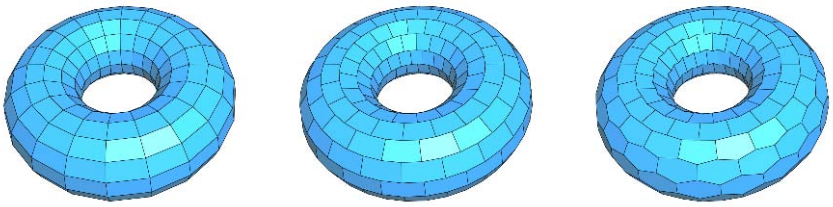


FIG. 8. *Left: a conjugate curve network; middle: the hexagonal mesh obtained by "shifting" alternative rows of the network in the left figure; right: a P-Hex mesh obtained by locally perturbing the hexagonal mesh in the middle figure.*

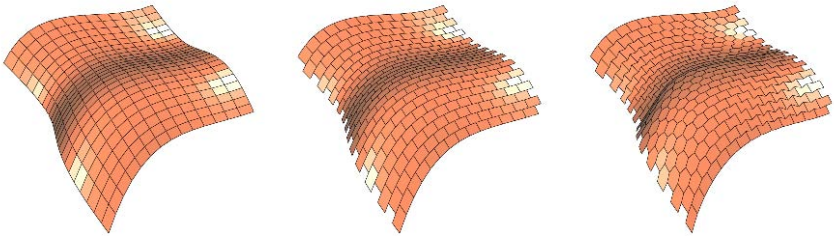


FIG. 9. *Computation of a P-Hex mesh approximating an open surface. Left: a conjugate curve network; middle: the intermediate hexagonal mesh; right: the final P-Hex mesh after local perturbation.*

layered supporting structures of a glass panel structure. The offset of a polyhedral surface is the discrete analogue of the offset surface of a smooth surface. There are several variants of the offset of a polyhedral surface; the most obvious one is the *constant-face-distance offset*, which is a polyhedral surface obtained by displacing each face of a given polyhedral surface by a constant distance along the normal of the face. In the following, a polyhedral surface will also be called a mesh, with the understanding that each face of the mesh is a planar polygon.

The offset mesh is closely related to the notion of *parallel meshes* — two meshes are *parallel* to each other if they are isomorphic and their corresponding edges have non-zero lengths and are parallel to each other. According to [6], the constant-distance offset of a smooth surface can be extended to the setting of polyhedral surfaces in three different ways: (1) constant face-distance offset; (2) constant edge-distance offset; and (3) constant vertex-distance offset. In terms of parallel meshes, a mesh M possesses a constant face-distance offset if it has a parallel mesh M' whose faces are tangent to S^2; a mesh M possesses a constant edge-distance offset if it has a parallel mesh M' whose edges are tangent to S^2; and a mesh M possesses a constant vertex-distance offset if it has a parallel mesh M'

whose vertices are on S^2. In the three cases above, the parallel mesh M' is called the *discrete Gaussian image* of the given mesh M. Then an offset mesh M_d with offset distance d of the mesh M is given by $M_d = M + d \cdot M'$, which is understood to be a vector expression for the corresponding vertices of the three meshes M, M' and M_d. We refer the reader to [6] for more detailed discussions about the definition, existence and construction of offset of general polyhedral surfaces.

In the following we will consider computing the offset meshes of P-Hex meshes. An equivalent condition for a mesh M with planar faces to possess a face-distance offset is that for every vertex v of M, all the faces incident to v are tangent to a common cone of revolution. For this reason, a mesh M possessing this property is also called a *conical mesh*. Obviously, this condition is satisfied by any P-Hex mesh, since there are exactly three faces incident to any vertex of a P-Hex mesh. That is, any P-Hex mesh is a conical mesh; as a consequence, any P-Hex mesh possesses constant face-distance offset P-Hex meshes.

Given a P-Hex mesh, its offset with the face-distance equal to a constant d can be computed as follows. For each vertex, we offset the three incident faces outward along their face normals by the distance d and intersect the three planes containing the three offset faces to determine the vertex of the offset mesh. Clearly, this approach will fail when the three faces are co-planar and it is numerically unstable when the three faces are nearly co-planar.

A more robust scheme is as follows. Let f_i, $i = 0, 1, 2$, be the three hex faces incident to a vertex v of a P-Hex mesh M. Let v_d be the vertex of the offset mesh M_d corresponding to v. Let N_i denote unit normal vectors of the f_i. Let θ_i be the internal angle of f_i at v. Then the "vertex" normal vector N_v of M ar v, defined by $v_d - v$, is parallel to

$$\bar{N}_v = \sum_{i=0}^{2} (\tan \beta_i + \tan \gamma_i) N_i,$$

where $\beta_i = \frac{1}{2}(\theta_i + \theta_{i+1} - \theta_{i-1})$ and $\gamma_i = \frac{1}{2}(\theta_i + \theta_{i-1} - \theta_{i+1})$, $i = 0, 1, 2$, mod 3. The proof of this formula is elementary so we omit it here. Using this formula, the vertex v_d can be determined by intersecting the line $p(t) = v + t\bar{N}_v$ with any one of the offset planes of the three faces. Figure 10 shows a P-Hex mesh with its constant face-distance offset mesh.

Next we consider the constant vertex-distance offset of a P-Hex mesh. An arbitrary P-Hex mesh may not possess a constant vertex-distance offset mesh. According to the above discussion, a necessary and sufficient condition for a P-Hex mesh M to have a constance vertex-distance offset is that it is parallel to a P-Hex mesh M' inscribed to the sphere S^2. Clearly, if there is such a mesh M', then every hex face of M' is inscribed to a circle. Let h' be a face of M'. Let α'_i, $i = 0, 1, \ldots, 5$, denote the six consecutively ordered internal angles of h' (see Figure 11). Then, since h' is inscribed to

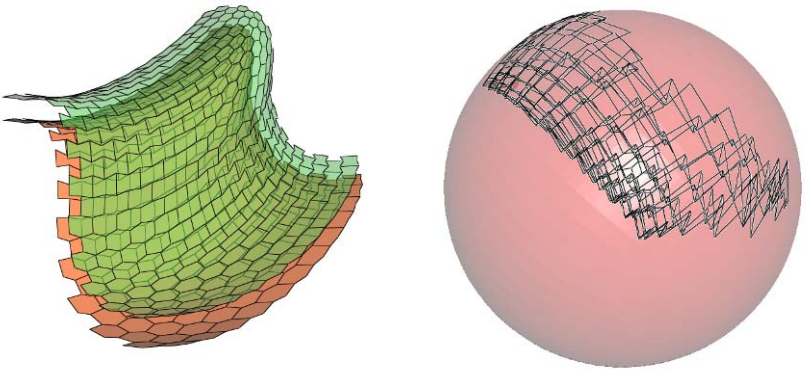


FIG. 10. *Left: a free-form P-Hex mesh and its constant face-distance offset mesh; right: the Gauss image of the P-Hex mesh, whose faces are tangent to the sphere* S^2.

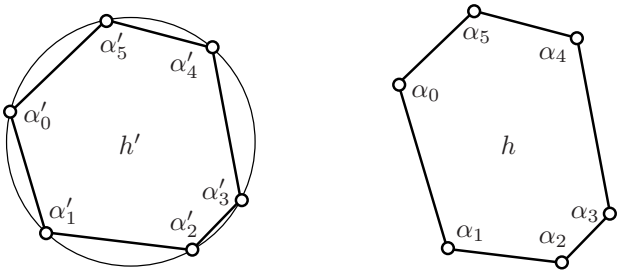


FIG. 11. *Left: a hex face* h' *inscribed to a circle; right: a hex face* h *parallel to* h'.

a circle, it is easy to show that $\alpha'_0 + \alpha'_2 + \alpha'_4 = \alpha'_1 + \alpha'_3 + \alpha'_5$. Let h be the hex face of M corresponding to h'. Let the α_i, $i = 0, 1, \ldots, 5$, be the corresponding angles of h. Since the edges of the face h are parallel to the edges of the face h' of M', $\alpha_i = \alpha'_i$. Therefore $\alpha_0 + \alpha_2 + \alpha_4 = \alpha_1 + \alpha_3 + \alpha_5$.

Conversely, it is easy to see that if

$$\alpha_0 + \alpha_2 + \alpha_4 = \alpha_1 + \alpha_3 + \alpha_5 \tag{4.1}$$

for a planar hex face h, then h is parallel to a hex face h' that is inscribed to a circle. From this it can be shown that, for an open P-Hex mesh M surface, it possesses a constant vertex distance offset mesh if and only if the angle condition (4.1) holds for every hex face h of M. That is to say, the angle condition (4.1) is a necessary and sufficient condition for a P-Hex mesh to possess constant vertex-distance offset meshes. Note that this angle condition (4.1) is only a necessary condition on this existence of

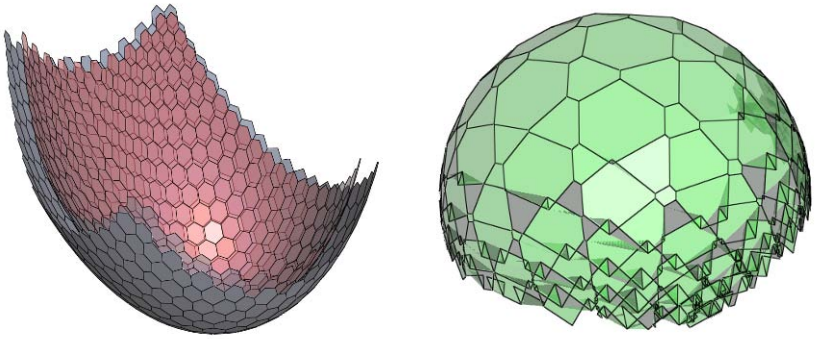


FIG. 12. *Left: a P-Hex satisfying the angle condition given by Eqn. (4.1), with its constant vertex-distance offset P-Hex mesh (superimposed); right: the Gauss image of the P-hex mesh, whose vertices are on the sphere* S^2.

constant vertex-distance offset meshes of a P-Hex mesh of a more complex topological type.

Figure 12 shows a P-Hex mesh whose faces satisfy the angle condition (4.1), together with its constant vertex-distance offset mesh, computed by integrating the angle condition (4.1) as a constraint in our local perturbation method.

According to [6], a P-Hex mesh possessing a constant edge-distance offset is necessarily parallel to a Koebe mesh, a mesh whose edges are tangent to the unit sphere S^2. This imposes significant restriction to the kind of surface shapes that can be represented by such P-Hex meshes. Also, the computation of the constant edge-distance offset meshes is more involved than the other types. For the detail we refer the reader to [6].

5. Further problems. There are numerous open problems calling for further research. First, it is important to understand the inherent degrees of freedom of P-Hex meshes tiling a free form surface. Such an understanding is fundamental to developing a general method for computing P-Hex meshes. Second, the issue of avoiding face self-intersection of P-Hex faces is still outstanding. All the existing methods, as well as the new method we have proposed here, cannot ensure that the computed P-Hex mesh is free of face self-intersection. We refer the reader to our recent technical report [8] on yet another method for generating P-Hex meshes based on tangent-duality and characterization of non-self-intersecting P-Hex faces in that context.

In view of practical applications in shape design, it would be desirable to be able to exert fine control over the the shape and size of the faces of a P-Hex mesh. Also, a subdivision scheme for P-Hex meshes would be very useful design tool. Finally, more research is needed on the computation

of offset meshes of P-Hex meshes, especially in the case of constant edge-distance offset meshes.

6. Acknowledgments. This research is supported by a General Research Fund (717808E) of Hong Kong Research Grant Council. The authors would like to thank Prof. Helmut Pottmann and Prof. Johannes Wallner for helpful discussions.

REFERENCES

[1] H. Almegaard, A. Bagger, J. Gravesen, B. Jüttler, and Z. Sir, *Surfaces with piecewise linear support functions over spherical triangulations*, in Proceedings of Mathematics of Surfaces XII, LNCS 4647, Springer, 2007.

[2] A. Bobenko, T. Hoffmann, and B. Springborn, *Minimal surfaces from circle patterns: Geometry from combinatorics*, Ann. of Math., **164**: 231–264, 2006.

[3] J. Díaz, C. Otero, R. Togores, and C. Manchado, *Power diagrams in the design of chordal space structures*, in The 2nd International Symposium on Voronoi Diagrams in Science and Engineering, pp. 93–104, 2005.

[4] H. Kawaharada and K. Sugihara, *Dual subdivision - a new class of subdivision schemes using projective duality*, in The 14th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision, pp. 9–17, 2006.

[5] Y. Liu, H. Pottmann, J. Wallner, Y. Yang, and W. Wang, *Geometric modeling with conical meshes and developable surfaces*, ACM Transactions on Graphics (SIGGRAPH 2006), **25**(3): 681–689, 2006.

[6] H. Pottmann, Y. Liu, J. Wallner, A. Bobenko, and W. Wang, *Geometry of multi-layer freeform structures for architecture*, ACM Transactions on Graphics (SIGGRAPH 2007), **26**(3), Article No. 65, 2007.

[7] D. Struik, *Lectures on classical differential geometry*, Cambridge, Addison-Wesley, 1950.

[8] W. Wang, Y. Liu, D. Yan, B. Chan, R. Ling, and F. Sun, *Hexagonal meshes with planar faces*, Technical Report, *TR-2008-13*, Department of Computer Science, The University of Hong Kong, 2008.

LIST OF WORKSHOP PARTICIPANTS

IMA Workshop on
Nonlinear Computational Geometry
May 29–June 2, 2007

- Iman Aganj, Department of Electrical and Computer Engineering, University of Minnesota
- Pankaj Kumar Agarwal, Department of Computer Science, Duke University
- Douglas N. Arnold, Institute for Mathematics and its Applications, University of Minnesota
- Donald G. Aronson, Institute for Mathematics and its Applications, University of Minnesota
- Xue Bai, Department Electrical and Computer Engineering, University of Minnesota
- Chandrajit L. Bajaj, Department of Computer Science, University of Texas
- Hélène Barcelo, Department of Mathematics and Statistics, Arizona State University
- Phillip Barry, Department of Computer Science and Engineering, University of Minnesota
- Saugata Basu, School of Mathematics, Georgia Institute of Technology
- Daniel J. Bates, Institute for Mathematics and its Applications, University of Minnesota
- Ciprian S. Borcea, Department of Mathematics, Rider University
- Laurent Busé Project GALAAD, Institut National de Recherche en Informatique Automatique (INRIA)
- Frédéric Cazals, Project Geometrica, Institut National de Recherche en Informatique Automatique (INRIA)
- Adarsh Chandran, Department Electrical and Computer Engineering, University of Minnesota
- Ionut Ciocan-Fontanine, Institute for Mathematics and its Applications, University of Minnesota
- David A. Cox, Department of Mathematics and Computer Science, Amherst College
- Carlos D'Andrea, Departament d'Algebra i Geometria, University of Barcelona
- Alicia Dickenstein, Departamento de Matematica - FCEyN, University of Buenos Aires
- Sandra Di Rocco, Department of Mathematics, Royal Institute of Technology (KTH)

I.Z. Emiris et al. (eds.), *Nonlinear Computational Geometry*, The IMA Volumes in Mathematics and its Applications 151, DOI 10.1007/978-1-4419-0999-2_10,

- Tor Dokken Department of Geometry SINTEF Kenneth R. Driessel, Department of Mathematics, Iowa State University
- Fayssal El Moufatich, Department of Intelligent Transportation Systems (VPE-ITS), Vodafone Group R&D Germany
- Ioannis Z. Emiris, Department of Informatics and Telecommunications, National Kapodistrian University of Athens
- Makan Fardad, Department of Electrical and Computer Engineering, University of Minnesota
- Michael S. Floater, Department of Informatics, University of Oslo
- Andre Galligo, Department of Mathematics, Université de Nice Sophia Antipolis
- Luis Garcia-Puente, Department of Mathematics, Texas A&M University
- Tryphon T. Georgiou, Department of Electrical Engineering, University of Minnesota
- Xavier Goaoc, Loria, Project Vegas, Institut National de Recherche en Informatique Automatique (INRIA)-Lorraine
- Ron Goldman, Department of Computer Science, Rice University
- Laureano Gonzalez-Vega, Departamento de Matemáticas, University of Cantabria
- Jason E. Gower, Institute for Mathematics and its Applications, University of Minnesota
- Milena Hering, Institute for Mathematics and its Applications, University of Minnesota
- Christopher Hillar, Department of Mathematics, Texas A&M University
- Benjamin J. Howard, Institute for Mathematics and its Applications, University of Minnesota
- Evelyne Hubert, Project CAFE, Institut National de Recherche en Informatique Automatique (INRIA)
- Manfred L. Husty, Unit Geometry and CAD, University Innsbruck
- Farhad Jafari, Department of Mathematics, University of Wyoming
- Ravi Janardan, Department of Computer Science and Engineering, University of Minnesota
- Itnuit Janovitz-Freireich, Department of Mathematics, North Carolina State University
- Anders Nedergaard Jensen, Institut for Matematiske Fag, Aarhus University
- Gabriela Jeronimo, Departamento de Matematica - FCEyN, University of Buenos Aires

- Steve Kaliszewski, Department of Mathematics and Statistics, Arizona State University
- Michael Kerber, Algorithms and Complexity, Max-Planck-Institut für Informatik
- Michael Kettner, School of Mathematics, Georgia Institute of Technology
- John Keyser, Department of Computer Science, Texas A&M University
- Rimvydas Krasauskas, Faculty of Mathematics and Informatics, Vilnius University
- Song-Hwa Kwon, Institute of Mathematics and its Applications, University of Minnesota
- Oliver Labs, Mathematik und Informatik, Universität des Saarlandes
- Niels Lauritzen, Institut for Matematiske Fag, Aarhus University
- Sylvain Lazard, Project Vegas, Institut National de Recherche en Informatique Automatique (INRIA)-Lorraine
- Anton Leykin, Institute for Mathematics and its Applications, University of Minnesota
- Hstau Y. Liao, Biochemistry and Molecular Biophysics, Columbia University
- Gennady Lyubeznik, School of Mathematics, University of Minnesota
- Dinesh Manocha, Department of Computer Science, University of North Carolina
- Hannah Markwig, Mathematisches Institut, Georg-August-Universität Göttingen
- Thomas Markwig, Department of Mathematics, Universität Kaiserslautern
- Kurt Mehlhorn,Max Planck Institute, Max-Planck-Institut für Informatik
- Richard B. Moeckel, School of Mathematics, University of Minnesota
- Brian Moore, Department of Symbolic Computation, Johann Radon Institute for Computational and Applied Mathematics
- Bernard Mourrain, Project GALAAD, Institut National de Recherche en Informatique Automatique (INRIA)
- Uwe Nagel, Department of Mathematics, University of Kentucky
- Jiawang Nie, Department of Mathematics, University of California, San Diego
- Dmitrii Pasechnik, School of Physical and Mathematical Sciences, Nanyang Technological University

- Martin Peternell, Institute of Discrete Mathematics and Geometry, Vienna University of Technology
- Jorg Peters, Computer Information Science and Engineering Department, University of Florida
- Ragni Piene, Centre of Mathematics for Applications, University of Oslo
- Mary Porter, Department of Genetics, Cell Biology, and Development, University of Minnesota
- Bharath Rangarajan, Department of Mechanical Engineering, University of Minnesota
- Tomas Recio, Departamento de Matemáticas, Estadística y Computación, University of Cantabria
- Victor Reiner, School of Mathematics, University of Minnesota
- Joel Roberts, School of Mathematics, University of Minnesota
- Marie Rognes, University of Oslo
- J. Maurice Rojas, Department of Mathematics, Texas A&M University
- Stergios I. Roumeliotis, Department of Computer Science and Engineering, University of Minnesota
- Bjarke Hammersholt Roune, Department of Mathematics, Aarhus University
- David Rusin, Department of Mathematical Sciences, Northern Illinois University
- Arnd Scheel,Institute for Mathematics and its Applications, University of Minnesota
- Chehrzad Shakiban, Institute of Mathematics and its Application, University of Minnesota
- Frank Sottile, Department of Mathematics, Texas A&M University
- Steven Sperber, School of Mathematics, University of Minnesota
- Reinhard Steffens, Department of Computer Science and Mathematics, Johann Wolfgang Goethe-Universität Frankfurt
- Ileana Streinu, Department of Computer Science, Smith College
- Agnes Szanto, Department of Mathematics, North Carolina State University
- Thorsten Theobald, Institut für Mathematik, Johann Wolfgang Goethe-Universität Frankfurt
- Louis Theran, Department of Computer Science, University of Massachusetts
- Carl Toews, Department of Mathematics, Duquesne University
- Nikolas Trawny, Department of Computer Science and Engineering, University of Minnesota

- Elias P. Tsigaridas, Projet VEGAS, Institut National de Recherche en Informatique Automatique (INRIA)
- John Voight, Department of Mathematics and Statistics, University of Vermont
- Wenping Wang, Department of Computer Science, University of Hong Kong
- Nicola Wolpert, Faculty of Geomatics, Computer Science and Mathematics, Stuttgart University of Applied Sciences
- William Wood, Department of Modeling and Simulation, Corning
- Fei Yang, Department of Biomedical Engineering, University of Minnesota
- Josephine Yu, Department of Mathematics, University of California
- Ming Zhang, Department of Biomathematical Sciences, University of Texas
- Xun Zhou, Department of Computer Science, University of Minnesota
- Severinas Zube, Department of Mathematics and Informatics, Vilnius State University

LIST OF WORKSHOP PARTICIPANTS

IMA Workshop on
Nonlinear Computational Geometry
May 29–June 2, 2007

- Iman Aganj, Department of Electrical and Computer Engineering, University of Minnesota
- Pankaj Kumar Agarwal, Department of Computer Science, Duke University
- Douglas N. Arnold, Institute for Mathematics and its Applications, University of Minnesota
- Donald G. Aronson, Institute for Mathematics and its Applications, University of Minnesota
- Xue Bai, Department Electrical and Computer Engineering, University of Minnesota
- Chandrajit L. Bajaj, Department of Computer Science, University of Texas
- Hélène Barcelo, Department of Mathematics and Statistics, Arizona State University
- Phillip Barry, Department of Computer Science and Engineering, University of Minnesota
- Saugata Basu, School of Mathematics, Georgia Institute of Technology
- Daniel J. Bates, Institute for Mathematics and its Applications, University of Minnesota
- Ciprian S. Borcea, Department of Mathematics, Rider University
- Laurent Busé Project GALAAD, Institut National de Recherche en Informatique Automatique (INRIA)
- Frédéric Cazals, Project Geometrica, Institut National de Recherche en Informatique Automatique (INRIA)
- Adarsh Chandran, Department Electrical and Computer Engineering, University of Minnesota
- Ionut Ciocan-Fontanine, Institute for Mathematics and its Applications, University of Minnesota
- David A. Cox, Department of Mathematics and Computer Science, Amherst College
- Carlos D'Andrea, Departament d'Algebra i Geometria, University of Barcelona
- Alicia Dickenstein, Departamento de Matematica - FCEyN, University of Buenos Aires
- Sandra Di Rocco, Department of Mathematics, Royal Institute of Technology (KTH)

I.Z. Emiris et al. (eds.), *Nonlinear Computational Geometry*, The IMA Volumes in Mathematics and its Applications 151, DOI 10.1007/978-1-4419-0999-2_10,

- Tor Dokken Department of Geometry SINTEF Kenneth R. Driessel, Department of Mathematics, Iowa State University
- Fayssal El Moufatich, Department of Intelligent Transportation Systems (VPE-ITS), Vodafone Group R&D Germany
- Ioannis Z. Emiris, Department of Informatics and Telecommunications, National Kapodistrian University of Athens
- Makan Fardad, Department of Electrical and Computer Engineering, University of Minnesota
- Michael S. Floater, Department of Informatics, University of Oslo
- Andre Galligo, Department of Mathematics, Université de Nice Sophia Antipolis
- Luis Garcia-Puente, Department of Mathematics, Texas A&M University
- Tryphon T. Georgiou, Department of Electrical Engineering, University of Minnesota
- Xavier Goaoc, Loria, Project Vegas, Institut National de Recherche en Informatique Automatique (INRIA)-Lorraine
- Ron Goldman, Department of Computer Science, Rice University
- Laureano Gonzalez-Vega, Departamento de Matemáticas, University of Cantabria
- Jason E. Gower, Institute for Mathematics and its Applications, University of Minnesota
- Milena Hering, Institute for Mathematics and its Applications, University of Minnesota
- Christopher Hillar, Department of Mathematics, Texas A&M University
- Benjamin J. Howard, Institute for Mathematics and its Applications, University of Minnesota
- Evelyne Hubert, Project CAFE, Institut National de Recherche en Informatique Automatique (INRIA)
- Manfred L. Husty, Unit Geometry and CAD, University Innsbruck
- Farhad Jafari, Department of Mathematics, University of Wyoming
- Ravi Janardan, Department of Computer Science and Engineering, University of Minnesota
- Itnuit Janovitz-Freireich, Department of Mathematics, North Carolina State University
- Anders Nedergaard Jensen, Institut for Matematiske Fag, Aarhus University
- Gabriela Jeronimo, Departamento de Matematica - FCEyN, University of Buenos Aires

- Steve Kaliszewski, Department of Mathematics and Statistics, Arizona State University
- Michael Kerber, Algorithms and Complexity, Max-Planck-Institut für Informatik
- Michael Kettner, School of Mathematics, Georgia Institute of Technology
- John Keyser, Department of Computer Science, Texas A&M University
- Rimvydas Krasauskas, Faculty of Mathematics and Informatics, Vilnius University
- Song-Hwa Kwon, Institute of Mathematics and its Applications, University of Minnesota
- Oliver Labs, Mathematik und Informatik, Universität des Saarlandes
- Niels Lauritzen, Institut for Matematiske Fag, Aarhus University
- Sylvain Lazard, Project Vegas, Institut National de Recherche en Informatique Automatique (INRIA)-Lorraine
- Anton Leykin, Institute for Mathematics and its Applications, University of Minnesota
- Hstau Y. Liao, Biochemistry and Molecular Biophysics, Columbia University
- Gennady Lyubeznik, School of Mathematics, University of Minnesota
- Dinesh Manocha, Department of Computer Science, University of North Carolina
- Hannah Markwig, Mathematisches Institut, Georg-August-Universität Göttingen
- Thomas Markwig, Department of Mathematics, Universität Kaiserslautern
- Kurt Mehlhorn,Max Planck Institute, Max-Planck-Institut für Informatik
- Richard B. Moeckel, School of Mathematics, University of Minnesota
- Brian Moore, Department of Symbolic Computation, Johann Radon Institute for Computational and Applied Mathematics
- Bernard Mourrain, Project GALAAD, Institut National de Recherche en Informatique Automatique (INRIA)
- Uwe Nagel, Department of Mathematics, University of Kentucky
- Jiawang Nie, Department of Mathematics, University of California, San Diego
- Dmitrii Pasechnik, School of Physical and Mathematical Sciences, Nanyang Technological University

- Martin Peternell, Institute of Discrete Mathematics and Geometry, Vienna University of Technology
- Jorg Peters, Computer Information Science and Engineering Department, University of Florida
- Ragni Piene, Centre of Mathematics for Applications, University of Oslo
- Mary Porter, Department of Genetics, Cell Biology, and Development, University of Minnesota
- Bharath Rangarajan, Department of Mechanical Engineering, University of Minnesota
- Tomas Recio, Departamento de Matemáticas, Estadística y Computación, University of Cantabria
- Victor Reiner, School of Mathematics, University of Minnesota
- Joel Roberts, School of Mathematics, University of Minnesota
- Marie Rognes, University of Oslo
- J. Maurice Rojas, Department of Mathematics, Texas A&M University
- Stergios I. Roumeliotis, Department of Computer Science and Engineering, University of Minnesota
- Bjarke Hammersholt Roune, Department of Mathematics, Aarhus University
- David Rusin, Department of Mathematical Sciences, Northern Illinois University
- Arnd Scheel,Institute for Mathematics and its Applications, University of Minnesota
- Chehrzad Shakiban, Institute of Mathematics and its Application, University of Minnesota
- Frank Sottile, Department of Mathematics, Texas A&M University
- Steven Sperber, School of Mathematics, University of Minnesota
- Reinhard Steffens, Department of Computer Science and Mathematics, Johann Wolfgang Goethe-Universität Frankfurt
- Ileana Streinu, Department of Computer Science, Smith College
- Agnes Szanto, Department of Mathematics, North Carolina State University
- Thorsten Theobald, Institut für Mathematik, Johann Wolfgang Goethe-Universität Frankfurt
- Louis Theran, Department of Computer Science, University of Massachusetts
- Carl Toews, Department of Mathematics, Duquesne University
- Nikolas Trawny, Department of Computer Science and Engineering, University of Minnesota

- Elias P. Tsigaridas, Projet VEGAS, Institut National de Recherche en Informatique Automatique (INRIA)
- John Voight, Department of Mathematics and Statistics, University of Vermont
- Wenping Wang, Department of Computer Science, University of Hong Kong
- Nicola Wolpert, Faculty of Geomatics, Computer Science and Mathematics, Stuttgart University of Applied Sciences
- William Wood, Department of Modeling and Simulation, Corning
- Fei Yang, Department of Biomedical Engineering, University of Minnesota
- Josephine Yu, Department of Mathematics, University of California
- Ming Zhang, Department of Biomathematical Sciences, University of Texas
- Xun Zhou, Department of Computer Science, University of Minnesota
- Severinas Zube, Department of Mathematics and Informatics, Vilnius State University