On Boundaries of Statistical Models

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

D I S S E R T A T I O N

zur Erlangung des akademischen Grades
DOCTOR RERUM NATURALIUM
(Dr.rer.nat.)

im Fachgebiet
Mathematik
vorgelegt

von Diplomphysiker Thomas Kahle
geboren am 17. Dezember 1981 in Magdeburg.

Die Annahme der Dissertation wurde empfohlen von:

1. Prof. Dr. Jürgen Jost (MPI-MiS, Leipzig)
2. Prof. Dr. Bernd Sturmfels (UC Berkeley, USA)

Die Verleihung des akademischen Grades erfolgt mit Bestehen
der Verteidigung am 26. Mai 2010 mit dem Gesamtprädikat
summa cum laude.

Contents

Acknowledgment

The research for this thesis was conducted in a very inspiring environment, making it most enjoyable. I thank my supervisor Nihat Ay for supporting my tremendous travel schedule, and for encouraging me to build my own research profile. Jürgen Jost, director of our institute, has supported me with many useful hints at just the right times. Bernd Sturmfels' visit to Germany in 2007/08 triggered my interest in algebraic statistics and algebra in general, and he has supported me in many ways. Johannes Rauh, doing his PhD on similar topics like mine, challenged me in uncountable discussions. We explored a lot of mathematics together. On my travels I had many inspiring discussions with Christian Haase, Raymond Hemmecke, Ezra Miller, Fero Matúš, Milan Studený, Seth Sullivant, and many others.

I thank my coauthors Nihat Ay, Nils Bertschinger, Jürgen Jost, Jürgen Herzog, Takayuki Hibi, Freyja Hreinsdóttir, Eckehard Olbrich, Johannes Rauh, and Walter Wenzel for conducting research together, and of course everybody who took the time to proofread the manuscripts of the thesis or the publications.

The inspiring atmosphere of the Max-Planck-Institute in Leipzig is made up by many factors. I want to explicitly mention Antje Vandenberg who made every little detail of organization run smoothly, and the fantastic library that has almost any book I ever wanted to look at, and instantly organized those they did not have. Personal thanks for so much crazy fun go in no particular order to: Wolfgang, Bastian, Bernhard, Rainer, Felix, Danijela, Guido, Jan, Eckehard, Keyan, Susanne, Pierre-Yves, Tobias, Stephan, Nils, ... and everybody else on jj_all@mis.

I want to thank my family for their continuous support. Finally, my deepest thanks go to Kai for her love, her creativity, and for teaching me the art of living.

The author's position was funded by the Volkswagen Foundation.

Introduction

Since the beginnings of statistics in the times of Sir Ronald Fischer, the problem of zero probabilities is a basic one. In this thesis we will focus on a specific instance of this problem. Consider the simplex of probability vectors on the set $\{1, \ldots, m\}$. A *statistical model* $\mathcal{M}$ is a "sufficiently nice" subset of the $(m-1)$-dimensional probability simplex inside $\mathbb{R}^m$. It is very natural to ask the following question about the model:

SUPPORT SET PROBLEM. *Describe the set of possible supports of* $\mathcal{M}$

$$\{S \subseteq \{1, \ldots, m\} : \exists\, p \in \mathcal{M} \text{ with } \operatorname{supp}(p) = S\}\,.$$

A priori, it is not at all clear how a solution to such a problem could look like. It might come in form of an oracle, which decides for a given set S, or as a quick algorithm that generates the list of sets. Both can be out of reach, but unresolvable questions can still serve as a motivation to move forward. This thesis summarizes and unifies the author's work on questions related to boundaries of statistical models [Kah10b; KWA09; RKA09] and algorithmic questions for binomial ideals [Kah10a]. The support set problem motivated many investigations that the author has undertaken during the research for this thesis. These investigations naturally also lead to other questions, but the support set problem is the leitmotif.

The use of geometry has lead to significant insights in many areas of mathematics, and there are two natural geometric ways to look at statistical models. One originates in differential geometry and studies local properties such as curvature [MR93; Ama00]. The other one is from the view point of algebraic geometry [DSS09; PRW01; PS05], whose techniques take a more global stance. In this thesis we are not so much concerned with differential geometry, but with implicit equations, placing it in the second field, called algebraic statistics.

It is in the nature of mathematics in general, and combinatorics in particular, that one strives to identify patterns. Nowadays, in principle, every iPod can perform algebraic geometry computations that needed a cluster computer 20 years ago. A typical work flow, employed by the author when studying a class of combinatorial objects, is to compute as many examples as possible with the help of pencil and computer, and then contemplate over the results until the general picture becomes visible. A prominent example are Markov bases. Computations carried out by the program 4ti2 [4ti207] lead to discovery of structure, and indeed, new theorems [LO06; Mal06; DS03; TA02]. It is therefore also in the scope of this work to show how computers help to find structures and present results of "experimental mathematics".

This thesis is divided into three chapters. The first deals with the boundaries of exponential families, also known as loglinear models. It lays the foundations of a combinatorial theory of exponential families that, although known to experts, was never fully documented in the literature. With the use of oriented matroid theory we develop implicit representations of exponential families, completely parallel to what is done in algebraic statistics, but without any assumptions on the sufficient statistics. Key results are the implicit representation of exponential families in Theorem 1.3.6

and a precise formulation of the correspondence between an exponential family and its associated polytope in Theorem 1.2.14.

In Chapter 2 hierarchical models are treated. The central concept is that of elementary circuits. Their lower support bound immediately gives a bound on Markov moves in Theorem 2.3.5 and the neighborliness of marginal polytopes in Theorem 2.5.2. Another result is Theorem 2.3.3, where we show simplicial complexes that give totally unimodular matrices and whose circuits are elementary. After that we discuss binary marginal polytopes, which are special in many ways. Among the results here is a classification of full-dimensional linear code polytopes in terms of their subgroups via the considerations leading to Theorem 2.4.12.

The last chapter is devoted to algorithms and software. We present *Binomials*, a software package with specialized algorithms for binomial ideals, including binomial primary decomposition. We discuss large primary decompositions that have been computed with the software. A central result and the conclusion of this thesis is Theorem 3.3.3, which gives a counterexample to conjectures of [ESU10].

To keep the treatment concise we will assume familiarity with basic notions from statistics and commutative algebra. Texts on statistics that have a geometric view are for instance [BN78] and [CS04]. A very pedagogical introduction to commutative algebra is [CLO96]; more advanced topics are covered extensively in [Eis95]. We will strive for brevity, not repeating material covered elsewhere if not absolutely necessary. Sections 1.3.1, 1.4.1, and 3.1.1 contain introductory material on realizable oriented matroids, toric varieties, and primary decomposition. After the bibliography an index is provided to ease the look-up of definitions. It also contains mathematical symbols.

CHAPTER 1

Discrete Exponential Families

1.1. Basic Notions

In this section we will be mostly concerned with definitions of basic objects. Let $[m] := \{1, \ldots, m\}$ be a nonempty finite set. Denote by $\mathbb{R}^m$ the vector space of real valued functions on $[m]$. We equivalently call $f \in \mathbb{R}^m$ a function or a vector and denote values, or components by $f(i)$, or f_i, respectively. Consider the closed set

$$\overline{\mathcal{P}_m} := \left\{ p \in \mathbb{R}^m : p(i) \geq 0 \, \forall i \in [m], \sum_{i=1}^{m} p(i) = 1 \right\} \tag{1.1}$$

of probability measures on $[m]$. It has the geometrical structure of a $(m-1)$-dimensional simplex whose extreme points are the unit vectors of $\mathbb{R}^m$. We often speak of the *simplex of probability distributions.* Every probability measure on a finite set is given by a function on the set itself instead of its power set. We speak of a probability distribution, or just a distribution. The overline already indicates that we use the notation $\overline{M}$ for the closure of $M \subseteq \mathbb{R}^m$ in the standard topology.

For every $f \in \mathbb{R}^m$ we denote

$$\operatorname{supp}(f) := \{i \in [m] : f(i) \neq 0\}, \tag{1.2}$$

its *support* and furthermore define

$$\mathcal{P}_m := \left\{ p \in \overline{\mathcal{P}_m} : \operatorname{supp}(p) = [m] \right\}, \tag{1.3}$$

the open simplex of distributions with full support. Throughout the text we use the convention $0 \log 0 = 0$, and log refers to the natural logarithm.

In this thesis we deal with random variables and variables in polynomial rings. We stick to the convention that no random variable is called simply a "variable", but always be preceded by the word "random".

It is very useful to have a geometric view on questions from information and probability theory. One step towards this is to introduce a notion of distance as follows:

Definition 1.1.1. Given $p, q \in \overline{\mathcal{P}_m}$ we call

$$D(p \parallel q) = \begin{cases} \sum_{i \in [m]} p(i) \log \frac{p(i)}{q(i)} & \text{if } \operatorname{supp}(q) \supseteq \operatorname{supp}(p) \\ \infty & \text{otherwise,} \end{cases} \tag{1.4}$$

the *Kullback Leibler distance of the distributions* p *and* q. It is also called *relative entropy* or *information divergence.*

This concept was introduced in [KL51] by Kullback and Leibler. They also considered a symmetric version $D(p \parallel q) + D(q \parallel p)$, which they called the *divergence.* Although not a metric, $D(p \parallel q)$ is nonnegative and zero if and only if $p = q$.

1.2. Exponential Families

One widely used class of statistical models is that of exponential families.

DEFINITION 1.2.1. Let $\mathcal{L} \subseteq \mathbb{R}^m$ be a linear subspace of $\mathbb{R}^m$. We call the set

$$\mathcal{E}_{\mathcal{L}} = \left\{ p \in \mathcal{P}_m : p(i) = \frac{\mathrm{e}^{H(i)}}{\sum_{j\in[m]} \mathrm{e}^{H(j)}}, H \in \mathcal{L} \right\} \tag{1.5}$$

the *exponential family* of the subspace $\mathcal{L}$ or the *loglinear model* of the subspace.

Due to the normalization, we can always assume that $\mathcal{L}$ contains the subspace spanned by the constant function $\mathbb{R}^m \ni \mathbb{1} : i \mapsto 1$, since including this subspace does not change the set $\mathcal{E}_{\mathcal{L}}$. Two functions H in (1.5) that differ by a constant lead to the same p. This also explains the loss of dimension in

EXAMPLE 1.2.2. The open $(m-1)$-dimensional probability simplex $\mathcal{P}_m$ is the exponential family of the whole, m-dimensional space $\mathbb{R}^m$.

Often the subspace in Definition 1.2.1 comes to us in a parametrized way. We are given a matrix $A = (a_{ij})_{i,j} \in \mathbb{R}^{d\times m}$ whose *rows* form a basis of $\mathcal{L}$. Otherwise we may choose a basis, and in the following any exponential family is given by a matrix A, such that we have a parameterization of $\mathcal{L}$:

$$\mathcal{L}_A := \left\{ c \cdot A : c \in \mathbb{R}^d \right\}. \tag{1.6}$$

The corresponding exponential family is denoted $\mathcal{E}_A$. The equivalent classical definition of an exponential family reads as

$$\mathcal{E}_A := \left\{ p \in \mathcal{P}_m : p(j) = \frac{1}{Z_c} \exp\left(\sum_{i=1}^d c_i a_{ij} \right),\ c \in \mathbb{R}^d \right\}, \tag{1.7}$$

where

$$Z_c := \sum_{j=1}^m \exp\left(\sum_{i=1}^d c_i a_{ij} \right). \tag{1.8}$$

Note that $\mathcal{E}_A$ is unchanged under row operations on the matrix A and additionally, because of the normalization, when adding a row $(1,\dots,1)$. We therefore assume that A has a row consisting of all entries 1. The linear map $A : \mathbb{R}^m \to \mathbb{R}^d$ is called the *sufficient statistics of* $\mathcal{E}_A$ for reasons that will become clear in Section 1.2.2.

EXAMPLE 1.2.3. Let $m = 4$, the probability simplex $\mathcal{P}_m$ is three-dimensional. We define an exponential family by the sufficient statistics

$$A := \begin{pmatrix} 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \end{pmatrix}. \tag{1.9}$$

The columns $a_1, \dots, a_4$ are indexed by $[m]$, the rows by the three parameters. The family is overparameterized. It is two-dimensional; the kernel of A is one-dimensional and spanned by the vector $(1,-1,-1,1)^T$. Describing the boundary points in parametrized form is not straightforward. Consider for instance the probability measure $p_{1,2} := (1/2, 1/2, 0, 0)^T$. We claim that $p_{1,2} \in \overline{\mathcal{E}_A}$. To see this define a sequence of parameter vectors $c_t := (t, 0, -t)$. Then $c_t A = (0, 0, -2t, -2t)$ and $p_t(j) = \frac{1}{Z_c} \exp(c_t A_j) \in \mathcal{E}_A$. It follows that as $t \to \infty$ we have $p_t \to p_{1,2}$ as desired. In this construction it is crucial to find a vector c that has the property that $cA_j = 0$ on the desired support and all signs of cA_j on the remaining columns agree. Note that the convex hull of the columns of A is a square lying in the plane defined by $x_3 = 1$. We found a vector c that does the job, only since the set $\{1, 2\}$ defines an edge of the square. This construction is not possible for the set $\{1, 4\}$, which defines a diagonal in the square. This connection between polyhedral geometry and support sets is at the heart of our analysis and will be an essential ingredient in the following.

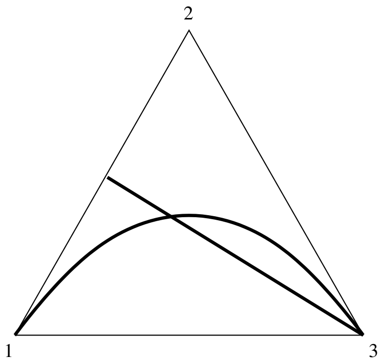


FIGURE 1. Degenerate exponential families

The following example illustrates other effects related to the relative positions of the columns of A.

EXAMPLE 1.2.4 (Families not containing all Dirac measures). Consider the (2×3)-matrix

$$A_1 = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 1 & 1 \end{pmatrix}. \tag{1.10}$$

The exponential model $\overline{\mathcal{E}_{A_1}}$ is one-dimensional and schematically depicted as the "curved" family in Figure 1. It does not contain all Dirac measures. More drastically, consider the family defined by

$$A_2 = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 1 & 1 \end{pmatrix}. \tag{1.11}$$

This family is also one-dimensional, but its closure contains only one Dirac measure, as depicted by the "straight" family in Figure 1. The sufficient statistics A_2 can not distinguish between the first and the second elementary event. Note that $\mathcal{E}_{A_2}$ is a convex exponential family, a rare case characterized in [MA03]. Constructions starting from a point configuration, like the ones we just considered, are often useful. For instance, in Example 5.7 of [Mat07] properties of the family can be read of from the picture. Note also how the geometry of $\mathcal{E}_A$ changes noncontinuously when perturbing the second column of A_2 by ϵ, which gives essentially another instance of the first example. Complementing these cases, in Example 1.3.17 we show an exponential family of dimension two, which contains all the Dirac measures.

REMARK 1.2.5 (Exponential families with base measures). The classical definition of an exponential family also allows the choice of a base measure q. Assuming a matrix A as above it is defined by

$$\mathcal{E}_{A,q} := \left\{ p \in \mathcal{P}_m : p(j) = \frac{q(j)}{Z_{c,q}} \exp\left(\sum_{i=1}^{d} c_i a_{ij} \right), c \in \mathbb{R}^d \right\}, \tag{1.12}$$

with

$$Z_{c,q} := \sum_{j=1}^{m} q(j) \exp\left(\sum_{i=1}^{d} c_i a_{ij}\right). \tag{1.13}$$

Evidently, if $\operatorname{supp}(q) = [m]$, then the support set problems solution is not effected by q, namely in any sequence of distributions, $p(i)$ goes to zero if and only if $\exp\left(\sum_{i=1}^{d} c_i a_{ij}\right)$ goes to zero. Note also, that *every* element of $\mathcal{E}_{A,q}$ can play the role of the base measure. See [CS04] for more details.

In contrast, if the support of q is not full, the family is restricted to a face of the probability simplex. Then one could argue that the choice of the set $[m]$ is not appropriate. The theory of experimental designs provides a possible approach to treat such models. This however always means a remodeling in terms of design matrices, briefly discussed in Section 1.4.3. For this thesis we will always assume an exponential family to be defined with full support as in (1.7).

1.2.1. Linear Families. Dual to the image of A, the kernel, its translations, and their intersections with the probability simplex, are also of importance. The linear space $\ker A$ can be "attached" to any probability measure, resulting in a linear family:

DEFINITION 1.2.6. The set

$$\mathcal{L}_{A,p} := (p + \ker A) \cap \overline{\mathcal{P}_m} := \left\{q \in \overline{\mathcal{P}_m} : Ap = Aq\right\}, \tag{1.14}$$

of probability vectors whose image under A equals Ap, is called a *linear family*.

As an intersection of an affine linear space with a polytope, $\mathcal{L}_{A,p}$ is itself a polytope. It is presented to us in its $\mathcal{H}$-representation, that is by inequalities. Naturally, it is an interesting, but nontrivial, question to determine the vertices of this polytope. We will next examine the relation between $\overline{\mathcal{E}_A}$ and $\mathcal{L}_{A,p}$.

1.2.2. Sufficient Statistics and Birch's Theorem. Loosely speaking, a *sufficient statistics* is a map of the data in some experiment to $\mathbb{R}^d$, such that the estimation of a distribution in the model can be carried out equally well, using either the data or the value of the sufficient statistics. More formally, let X be a random variable distributed according to a fixed and unknown $p \in \mathcal{P}(\mathcal{X})$, from which we obtain r samples $u = (u^{(1)}, \dots, u^{(r)})$, and U a random variable that consists of r independent identical copies of X. Then a map T is called a sufficient statistics for a model $\mathcal{E}$, if as long as the unknown distribution p of X is contained in $\mathcal{E}$, the conditional probability $\mathbb{P}(U = u | T(U) = T(u))$ is independent of P. This fact is commonly expressed as "Given the sufficient statistics, the conditional probability of observing the data is independent of the parameters". It implies that the value of the sufficient statistics carries all information that can be used in inference.

For the exponential models $\mathcal{E}_A$, defined in the last section, it will instantly turn out that the matrix A itself computes a sufficient statistics. The following theorem is classical.

THEOREM 1.2.7 (Birch's Theorem). *Let $A \in \mathbb{R}^{d\times m}$ and $p \in \overline{\mathcal{P}_m}$ a probability vector. Then there is a unique point in the intersection of the associated linear and exponential family:*

$$\{p^*\} := \mathcal{L}_{A,p} \cap \overline{\mathcal{E}_A}. \tag{1.15}$$

The distribution p^* *is called the* maximum likelihood estimate (MLE) *of* p *in the model* $\overline{\mathcal{E}_A}$. *Equivalently* p^* *is the unique solution to the* likelihood equations*:*

$$Ap^* = Ap, \qquad p^* \in \overline{\mathcal{E}_A}. \tag{1.16}$$

A very nice geometric proof of this fact can be found in the considerations leading to Theorem 3.3 of [CS04]. More generally, the *Pythagorean identity* is shown. With the notation from above, for any $q \in \overline{\mathcal{E}_A}$ one has the following identity of Kullback-Leibler distances:

$$D(p \parallel q) = D(p \parallel p^*) + D(p^* \parallel q). \tag{1.17}$$

This equation can also be used to reformulate the problem of computing the solution, i.e. finding maximum likelihood estimates, as a minimization of $D(p \parallel q)$. This is carried out either in the first argument over the polytope $\mathcal{L}_{A,p}$, or in the second argument over the models closure $\overline{\mathcal{E}_A}$.

Remark 1.2.8 (Finding the solution of (1.16)). Owing to the constraint $p^* \in \overline{\mathcal{E}_A}$, it is not generally possible to solve the likelihood equations (1.16) in closed form. In Section 1.3 it will be shown that the requirement $p^* \in \overline{\mathcal{E}_A}$ is equivalent to certain equations connected to the oriented matroid of the matrix A. Under additional conditions on A, the requirement is equivalent to *polynomial* equations, and the full set of likelihood equations is actually a polynomial system. Geometrically, Birch's Theorem then states that, inside the probability simplex, a linear space and a toric variety intersect in a unique nonnegative real point. The solution can be found numerically, using a simple algorithm called *iterative proportional fitting (IPF)*, described for instance in [Stu02]. An open source implementation is the software CIPI [SK08].

A point $p^* \in \overline{\mathcal{E}_A}$ is called the maximum likelihood estimate (MLE) of $p \in \overline{\mathcal{P}_m}$ if it is the solution to the problem (1.15). In statistics literature only a point in the interior of $\mathcal{E}_A$ is called an MLE. If the solution lies on the boundary, the MLE is said to not exist. This notion originates in the fact that exponential families are mostly considered as parameterized models and the classical parameterizations, generated by parameterizing the linear space in (1.5), do not extend to the boundary. It is clearly visible how useful a unified treatment of boundary and interior points is. However, already defining the closure of general exponential families for continuous random variables is a highly nontrivial problem [CM04; CM05]. In our discrete case the notions considered in these papers all coincide with the usual closure taken in $\mathbb{R}^d$; the problems lie more in finding explicit descriptions of elements in the boundary. In the next section we will show the fundamental role of the matrix A in such problems.

1.2.3. Polytopes of Exponential Families. Given an exponential family $\mathcal{E}_A$, it has been visible in the examples that the convex hull of the columns of A plays a certain role, which we discuss now. We denote the columns of A as $\{a_1, \ldots, a_m\}$. Note that these vectors are the values that the sufficient statistics takes on the Dirac measures, the extreme points of the probability simplex. We consider the polytope of possible values that Ap can take for $p \in \overline{\mathcal{E}_A}$. As a map on the probability simplex, A is not injective. However, the preimages of points p are just the linear families $\mathcal{L}_{A,p}$, which intersect with $\overline{\mathcal{E}_A}$ in a unique point. Thus, each point in the simplex has a corresponding point in $\overline{\mathcal{E}_A}$ with the same value under A, and we have a linear foliation of the simplex as

$$\overline{\mathcal{P}_m} = \dot{\bigcup_{p \in \overline{\mathcal{E}_A}}} \mathcal{L}_{A,p}. \tag{1.18}$$

Since every probability vector is a convex combination of Dirac measures and A is linear, the image of the exponential family $\overline{\mathcal{E}_A}$ under A is a polytope, given by the convex hull of the columns of A. Up to now, as we only assumed a row of ones, essentially any polytope can occur. Nevertheless we assign a name:

DEFINITION 1.2.9. The convex hull of the columns of A

$$\operatorname{cs}(\overline{\mathcal{E}_A}) := \operatorname{conv}\{a_j, j = 1, \ldots, m\} \tag{1.19}$$

is called the *convex support* of the exponential family.

REMARK 1.2.10. A common definition reads as follows: Let μ be a Borel measure on $\mathbb{R}^d$. The convex support is the intersection of all closed convex sets C, with full measure, $\mu(C) = \mu(\mathbb{R}^d)$. In this sense each $p \in \mathcal{E}_A$ induces a (finitely supported) measure on $\mathbb{R}^d$ via $\mu(C) := p\left(A^{-1}(C)\right)$, where we have slightly abused notation by denoting $A^{-1}(C)$ the preimage of C under A. Then the convex hull of the columns of A is the convex support of every element $p \in \mathcal{E}_A$.

The convex support plays a central role in the first part of this thesis. It encodes the solution of the support set problem in its face lattice. The remaining parts of this chapter are to clarify the connection between $\operatorname{cs}(\overline{\mathcal{E}_A})$ and $\overline{\mathcal{E}_A}$. The following three theorems justify $\operatorname{cs}(\overline{\mathcal{E}_A})$ to be called the combinatorial, or convex, version of $\overline{\mathcal{E}_A}$. We start by considering only the topological properties of $\operatorname{cs}(\overline{\mathcal{E}_A})$ and $\overline{\mathcal{E}_A}$.

THEOREM 1.2.11. *The closure $\overline{\mathcal{E}_A}$ of a discrete exponential family $\mathcal{E}_A$ is homeomorphic to its convex support* $\operatorname{cs}(\overline{\mathcal{E}_A})$.

PROOF. The proof is given by explicitly studying the maps between these two objects. The only difficulty is to prove Birch's Theorem 1.2.7. The sufficient statistics restricted to the probability simplex

$$A : \overline{\mathcal{P}_m} \to \mathbb{R}^d, \tag{1.20}$$

is continuous being the restriction of a linear map. The opposite map is a bit harder to understand. For each $\mu \in \operatorname{cs}(\overline{\mathcal{E}_A})$ we define $\phi(\mu)$ to be the unique point $p \in \overline{\mathcal{E}_A}$ with $Ap = \mu$, which exists by Birch's Theorem 1.2.7. This is clearly a bijection. Now, since A is a continuous map from a compact set to a compact set, it is a closed map, and the inverse is also continuous. $\square$

REMARK 1.2.12. Variants of the theorem are known for very long time. In the interior of the polytope and the family, the statement is a statement about the existence and uniqueness of maximum likelihood estimates, a weaker variant of Birch's Theorem. The critical part here is the extension to the boundary and the preservation of combinatorial structure (Theorems 1.2.13 and 1.2.14). Authors often speak of the homeomorphism of an exponential family and the associated polytope, but this is not enough; a homeomorphism onto a polytope is not better than a homeomorphism onto a ball of correct dimension. It is not straightforward to state the theorem as a homomorphism of two objects. We first state this fact as it is in Theorem 1.2.14 and then use more fancy names in Theorem 1.2.18, where we introduce the notion of *regular cell complexes.* If the exponential family is a toric variety, a case that will be discussed below, the homeomorphism has been called the *moment map* by authors from algebraic geometry.

THEOREM 1.2.13. *The homeomorphism in Theorem 1.2.11 respects the boundary in the following way: A set $F \subseteq [m]$ is the support of an element $p \in \overline{\mathcal{E}_A}$, if and only if, it is the preimage (under the map induced by A) of a set of vertices of a face of* $\operatorname{cs}(\overline{\mathcal{E}_A})$.

This theorem can be deduced from general statements about closures of exponential families [CM05; CM03]. Our case of discrete random variables is already contained in Chapter 9 of [BN78]. Some more discussion and pointers to the literature can be found in [KA06]. For this work we do not need the full generality and explicitly prove a slightly weaker version, which excludes the effects discussed in Example 1.2.4.

THEOREM 1.2.14. *Assume that the columns of A are the vertices of their convex hull. If $p \in \overline{\mathcal{E}_A}$ with $F = \operatorname{supp}(p)$, then $\operatorname{conv}\{a_i : i \in F\}$ is a face of $\operatorname{cs}(\overline{\mathcal{E}_A})$, containing Ap. Conversely if $\mu \in \operatorname{cs}(\overline{\mathcal{E}_A})$ lies in the relative interior of a face defined by $F \subseteq [m]$, then the unique preimage $\hat{p}$ satisfies $\operatorname{supp}(\hat{p}) = F$.*

Our proof of Theorem 1.2.14 will use the implicit representation of exponential families, to be introduced in the next section. Its main part consists of Theorem 1.3.1 and Lemma 1.3.4, but we already give the theorem here to sum up the considerations about the sufficient statistics. The statement will not be used in the considerations of Section 1.3 and the proof is given right before Proposition 1.3.5 on page 13.

EXAMPLE 1.2.15. In Example 1.2.4, the convex hull of the columns of A_2 is a line segment. The preimage of its "left vertex" is the set $\{1, 2\}$, which supports a probability measure in $\overline{\mathcal{E}_{A_2}}$. However, Theorem 1.2.14 fails: $\operatorname{conv}\{a_1\} = \{a_1\}$ is a face of the convex support while it does not give the support of a probability measure in $\overline{\mathcal{E}_{A_2}}$.

We continue with a reformulation of the homeomorphism with preservation of boundary structure in a more abstract fashion, using regular cell complexes.

DEFINITION 1.2.16. A *regular cell complex* Δ is a collection of closed topological balls σ in a Hausdorff space such that, denoting $\|\Delta\| := \bigcup_{\sigma\in\Delta} \sigma$, we have

(1) The interiors σ° partition Δ, i.e. every $x \in \|\Delta\|$ lies in exactly one σ°.
(2) The boundary $\partial\sigma$ is a union of members of Δ for any $\sigma \in \Delta$.

The elements $\sigma \in \Delta$ are called *closed cells*, their interiors σ° *open cells*, and $\|\Delta\|$ is called the *underlying space* of the cell complex. If some topological space T satisfies $T \cong \|\Delta\|$, then Δ is said to provide a *regular cell decomposition* of T. The *face poset* of Δ is the set of closed cells ordered by inclusion. Two regular cell complexes Δ, Γ are called *isomorphic* if there exists a homeomorphism $\phi : \|\Delta\| \to \|\Gamma\|$ such that the restriction ϕ_σ, of ϕ to any cell $\sigma \in \Delta$, is a homeomorphism of σ and a cell of Γ.

With this definition any polytope is a regular cell complex. Its face poset is just the lattice of faces in the sense of polyhedral geometry. In particular the probability simplex is a cell complex and induces a cell decomposition of $\overline{\mathcal{E}_A}$:

PROPOSITION 1.2.17. *The closed exponential family $\overline{\mathcal{E}_A}$ admits a cell decomposition where the open cells are the intersections*

$$\left\{\overline{\mathcal{E}_A} \cap F^\circ : F \subseteq \overline{\mathcal{P}_m} \text{ a face}\right\}. \tag{1.21}$$

PROOF. The elements p in nonempty intersections $\overline{\mathcal{E}_A} \cap F^\circ$ consist exactly of those distributions with $\operatorname{supp}(p) = F$. Thus they partition $\overline{\mathcal{E}_A}$ according to the possible support sets. The boundary of a cell σ consists of those distributions in $\overline{\mathcal{E}_A}$ whose support is contained in the support of elements in σ, and these all occur among the intersections in (1.21). $\square$

Finally, this allows a reformulation of Theorem 1.2.14 as:

THEOREM 1.2.18. *The map $A : \overline{\mathcal{E}_A} \to \operatorname{cs}(\overline{\mathcal{E}_A})$ is an isomorphism of regular cell complexes.*

1.3. Implicit Representations of Exponential Families

Let us summarize our current knowledge about the boundary of discrete exponential families. The exponential family is mapped to a convex polytope by the restriction of a linear map. This map is in fact a bijection on the model $\overline{\mathcal{E}_A}$ by Theorem 1.2.11. The boundary and thereby the possible support sets are characterized by the face lattice of the convex support, as stated in Theorem 1.2.14. To prove it, we will introduce a very useful tool, namely implicit representations of exponential families. Using parameterizations to describe the boundary of a statistical model is troublesome. A much more powerful and successful technique is to use implicitization, which means to describe the exponential family as the solutions to a set of equations. When it comes to proofs in the derivation of such a representation one naturally encounters the structure of the oriented matroid of the columns of the sufficient statistics A. After having discussed the implicit representations, we will give a short introduction into this theory and show how it relates to what we have presented here.

Apart from describing exponential families implicitly, our goal in this section, is to find a concise characterization of the support sets in general exponential families with the help of oriented matroids. Although slightly hidden, the connection to oriented matroid theory is natural. The presentation here is largely inspired by Markov bases, which provide an implicit description of exponential families for discrete random variables with integer valued sufficient statistics A. When one leaves the special case of commutative algebra and toric ideals, what remains is oriented matroid theory. The Implicitization Theorem 1.3.1 parallels the celebrated representation of toric statistical models. Here, we study the–not necessarily polynomial–equations that define the closure of the exponential family and relate them to the oriented matroid of the sufficient statistics of the model. In the toric case, our observations reduce to the fact that the nonnegative real part of a toric variety is described by a circuit ideal. We emphasize how the proof of this fact uses arguments from oriented matroid theory. While encountering only realizable oriented matroids, we give pointers to generalizations. Textbook references on the subject are [BVS$^+$93] and a chapter in [Zie94].

The first theorem shows how to obtain an implicit description of $\overline{\mathcal{E}_A}$ from the kernel of A. Note, how this gives a nice "duality" as the parametrization itself is derived from the image of A. For any given vector $n \in \mathbb{R}^m$, we denote n^+ and n^- its positive, respectively negative, part with components $n^+(x) := \max(0, n(x))$ and $n^-(x) := \max(0, -n(x))$, which gives $n = n^+ - n^-$. To make the following equations concise, we introduce *monomial notation*, where for any $n \in \mathbb{N}^m$ we denote $p^n := \prod_{i=1}^m p(i)^{n(i)}$.

Theorem 1.3.1. *A distribution p is an element of the closure $\overline{\mathcal{E}_A}$ if and only if it satisfies all the equations*

$$p^{n^+} = p^{n^-}, \qquad \text{for all } n \in \ker A. \tag{1.22}$$

Remark 1.3.2. This theorem is a direct generalization of Theorem 3.2 in [GMS06]. There only the polynomial equations among (1.22) are studied under the additional assumption that A has only integer entries. However, the proof of the theorem generalizes without any major problem. Actually, the proof of our theorem needs one step less, since we don't need to show the reduction to the polynomial equations. The different flavor of the results will be made more precise in Section 1.4. It is easily seen that an analogous statement holds for an arbitrary base measure with full support. In [RKA09] the general statement is given. Our proof closely follows [GMS06], however, we explicitly point out how matroid-type arguments are used, the first example being Lemma 1.3.4.

Before giving the proof of Theorem 1.3.1 we first state a couple of auxiliary results that are of independent interest. The matrix A and derived objects are fixed for the rest of the considerations. A *face* of a polytope P is the intersection of the polytope with an affine hyperplane H, such that all $x \in P$ with $x \notin H$ lie on one side of the hyperplane. Faces of maximal dimension are called *facets*. It is a fundamental result that every polytope can equivalently be described as a compact set defined by finitely many inequalities (i.e. facets), see [Zie94].

In particular we are interested in the face structure of $\mathrm{cs}(\overline{\mathcal{E}_A})$, as this gives us the set of possible supports of the model $\overline{\mathcal{E}_A}$. Our assumption that A has a row $(1,\dots,1)$ in this setting means that all vertices of $\mathrm{cs}(\overline{\mathcal{E}_A})$ lie in the affine hyperplane $x_d = 1$. Thus, for any face, we can replace the affine hyperplane H by an equivalent central hyperplane, passing through the origin. This motivates the following

Definition 1.3.3. Let $\{a_i : i \in [m]\}$ be the vertex set of a polytope, lying in a hyperplane that does not contain zero. A set $F \subseteq [m]$ is called *facial* if there exists a vector $c \in \mathbb{R}^d$ such that

$$c^T a_i = 0 \quad \forall i \in F, \qquad\qquad c^T a_j \geq 1 \quad \forall j \notin F. \tag{1.23}$$

Lemma 1.3.4. *Fix a matrix $A = (a_i)_{i\in[m]} \in \mathbb{R}^{d\times m}$ and a nonempty subset $F \subseteq [m]$. Then we have:*

- *If F is facial then no nonzero nonnegative linear combination of the a_i, $i \notin F$, can be written as linear combination of the a_i, $i \in F$.*
- *F is facial if and only if for any $u \in \ker A$:*

$$\mathrm{supp}(u^+) \subseteq F \Leftrightarrow \mathrm{supp}(u^-) \subseteq F. \tag{1.24}$$

- *If p is a solution to* (1.22), *then* $\mathrm{supp}(p)$ *is facial.*

The proofs here use standard arguments from oriented matroid theory and polytope theory. The last part is following the argumentation in the appendix of [GMS06]. We reproduce it to show some oriented matroid theory in everyday use.

Proof. For the first statement, assume to the contrary that we can find $\alpha(i) \geq 0$ and $\beta(i)$ not all zero such that $u = \sum_{i\notin F} \alpha(i) a_i = \sum_{i\in F} \beta(i) a_i$, and let c be normal to the facial hyperplane as in (1.23). We have

$$0 \leq \sum_{i\notin F} \alpha_i \leq \sum_{i\notin F} \alpha_i c^T a_i = c^T \left(\sum_{i\notin F} \alpha_i a_i \right) = c^T \left(\sum_{i\in F} \beta_i a_i \right) = 0, \tag{1.25}$$

whence $\alpha_i = 0$ for all $i \notin F$. Still under the assumption that F is facial and additionally $u = u^+ - u^- \in \ker A$, we see that $\mathrm{supp}(u^+) \subseteq F \Leftrightarrow \mathrm{supp}(u^-) \subseteq F$.

For the opposite direction we use Farkas' Lemma (see for example [Zie94]): *Let $B \in \mathbb{R}^{l\times d}$, and $z \in \mathbb{R}^l$. Either there exists a point in the polyhedron $\{x : Bx \leq z\}$, or there exists a nonnegative vector $y \in \mathbb{R}^l_{\geq}$ with $y^T B = 0$ and $y^T z < 0$, but not both.* Assume that $F \subsetneq [m]$ is nonempty and satisfies (1.24) for all $u \in \ker A$. Let B be the $(|F| + m) \times d$ matrix with rows $\{a_i^T : i \in F\}$, $\{-a_i^T : i \in F\}$, $\{-a_i^T, i \notin F\}$, and z be the vector which has entries zero in the first $2|F|$ components and entries -1 in the last $m - |F|$. Then a solution to $Bx \leq z$ provides a facial vector. Thus it remains to show that each nonnegative $y = (y^{(1)}, y^{(2)}, y^{(3)})^T$, decomposed according to the rows of B, with $y^T B = 0$ satisfies $y^T z \geq 0$. Without loss of generality, assume that the columns of A are ordered such that the columns with indices $i \in F$ come first. Then $y^{(3)}$ must be zero as otherwise $(y^{(2)} - y^{(1)}, y^{(3)})^T \in \ker A$ would violate (1.24) by nonnegativity of y. In this case $y^T z = 0$ holds trivially.

The last statement uses a simple argument about the supports of exponents versus the supports of distributions, which we use from time to time. Assume that p satisfies (1.22) and let $u = u^+ - u^- \in \ker A$. We show that

$$\operatorname{supp}(u^+) \subseteq \operatorname{supp}(p) \Leftrightarrow \operatorname{supp}(u^-) \subseteq \operatorname{supp}(p). \tag{1.26}$$

Assume that $i \in \operatorname{supp}(u^+)$ with $i \notin \operatorname{supp}(p)$. Then one, and thus both sides of (1.22) equal zero. Evidently there exist $j \in \operatorname{supp}(u^-)$ with $j \notin \operatorname{supp}(p)$. The same argument applies for the other direction and the second statement shows that $\operatorname{supp}(p)$ is facial. □

Now we are ready for the proof of Theorem 1.3.1.

Proof of Theorem 1.3.1. Denote $Z_A \subseteq \overline{\mathcal{P}_m}$ the set of solutions of (1.22). We first show that $\mathcal{E}_A$ satisfies the equations defining Z_A. Let $p \in \mathcal{E}_A$, using the parameterization we can write $p(i) = \mathrm{e}^{\theta^T a_i}$, for some vector of parameters $\theta \in \mathbb{R}^d$ and a_i the i-column of A. We find

$$\begin{aligned} p^u = \prod_{i\in[m]} p(i)^{u(i)} &= \prod_{i\in[m]} \left(\mathrm{e}^{\theta^T a_i}\right)^{u(i)} \\ &= \prod_{i\in[m]} \mathrm{e}^{\theta(i)(Au)(i)} = \prod_{i\in[m]} \mathrm{e}^{\theta(i)(Av)(i)} = p^v, \end{aligned} \tag{1.27}$$

for each $u, v \in \mathbb{N}^{[m]}$ with $Au = Av$. Thus $\mathcal{E}_A \subseteq Z_A$, and also $\overline{\mathcal{E}_A} \subseteq \overline{Z}_A = Z_A$.

Next, let $p \in Z_A \setminus \mathcal{E}_A$. We show that p is the limit of distributions in $\mathcal{E}_A$ by constructing a sequence p_μ in $\mathcal{E}_A$ that converges to p as $\mu \to -\infty$. Consider the following system of equations in variables $b = (b_1, \dots, b_d)$:

$$b^T a_i = \log p(i) \qquad \text{for all } i \in \operatorname{supp}(p). \tag{1.28}$$

We claim that this linear system has a solution. Denote $A_F \in \mathbb{R}^{d\times|F|}$ the matrix that has the columns $\{a_i : i \in F\}$. Then if (1.28) has no solution, the right hand side is not in the row space of A_F and there exists a vector $q \in \ker A_F$ with $\sum_{j\in F} q(j) \log p(j) \neq 0$. We extend q to the kernel of A by defining

$$u(i) := \begin{cases} q(i) & \text{if } i \in F, \\ 0 & \text{otherwise,} \end{cases} \tag{1.29}$$

and write $u = u^+ - u^-$. It follows that $u \in \ker A$ since $Au = A_F q$, on the other hand we have

$$\sum_{j\in F} u^+(j) \log p(j) \neq \sum_{j\in F} u^-(j) \log p(j), \tag{1.30}$$

and

$$\prod_{j\in[m]} p(j)^{u^+(j)} \neq \prod_{j\in[m]} p(j)^{u^-(j)}, \tag{1.31}$$

but equality holds since $p \in Z_A$. From this contradiction we deduce that (1.28) has a solution. Now, fix a vector $c \in \mathbb{R}^d$ with property (1.23) and for any $\mu \in \mathbb{R}$ define

$$p_\mu := \left(\mathrm{e}^{\mu c^T a_1 + b^T a_1}, \dots, \mathrm{e}^{\mu c^T a_m + b^T a_m}\right) \in \mathcal{E}_A. \tag{1.32}$$

By (1.23) it is clear that $\lim_{\mu\to-\infty} p_\mu = p$. This proves the theorem. □

The construction of the sequence in the proof was already used in Example 1.2.3. With Theorem 1.3.1 we are able to settle Theorem 1.2.14:

PROOF OF THEOREM 1.2.14. For the first statement, let $p \in \overline{\mathcal{E}_A}$ with $\operatorname{supp}(p) = F$ be given. It satisfies the equations (1.22) and thus F is facial. Note also that p can be written uniquely in terms of Dirac measures $p = \sum_{i \in F} p(i)\delta_i$. Thus $Ap \in \operatorname{conv}\{a_i : i \in F\}$.

For the second statement let $\mu = \sum_{i\in F} \lambda_i a_i$ and denote by $\hat{p}$ be the unique MLE of μ in $\overline{\mathcal{E}_A}$. From

$$A\hat{p} = \sum_{i\in F} \hat{p}(i)a_i + \sum_{j\notin F} \hat{p}(j)a_j = \mu = \sum_{i\in F} \lambda_i a_i, \tag{1.33}$$

we see

$$\sum_{j\notin F} \hat{p}(j)a_j = \sum_{i\in F} \lambda_i a_i - \sum_{i\in F} \hat{p}(i)a_i, \tag{1.34}$$

and the first statement of Lemma 1.3.4 shows that $\hat{p}(j) = 0, j \notin F$ and $\operatorname{supp}(\hat{p}) \subseteq F$. Now, if $\operatorname{supp}(\hat{p}) = G \subsetneq F$, then G is also facial and contains $\mu = A\hat{p}$. This contradicts the fact that μ lies in the relative interior of the face defined by F. It follows that $\operatorname{supp}(\hat{p}) = F$. □

In [GMS06] a stronger version of the last statement of Lemma 1.3.4 is given. This now follows from Theorem 1.2.14.

PROPOSITION 1.3.5. *The following are equivalent for any set $F \subseteq [m]$:*

(1) F is facial.
(2) The uniform distribution $\frac{1}{|F|}\mathbb{1}_F$ of F lies in $\overline{\mathcal{E}_A}$.
(3) There is a vector with support F in $\overline{\mathcal{E}_A}$.

PROOF. When F is facial, any element $u \in \ker A$ satisfies (1.24). This guarantees that the uniform distribution satisfies the equations (1.22) and thus is contained in $\overline{\mathcal{E}_A}$. Therefore (1) $\Rightarrow$ (2). The implication (2) $\Rightarrow$ (3) is clear and (3) $\Rightarrow$ (1) follows from Lemma 1.3.4. □

According to Theorem 1.3.1 the set $\overline{\mathcal{E}_A}$ is characterized by infinitely many equations. We now reduce them to finitely many. If these equations were actually polynomial, they would define an ideal and the Hilbert Basis Theorem would ensure that finitely many equations suffice. As we do not consider this assumption here we take a different route by studying the combinatorial essence of the matrix A, its oriented matroid. For this, we need the following notion from matroid theory: A *circuit vector* of a matrix A is a nonzero vector $n \in \mathbb{R}^m$ corresponding to a linear dependency $\sum_i n(i)a_i$ with inclusion minimal support, i.e. if $n' \in \mathbb{R}^m$ satisfies $\operatorname{supp}(n') \subseteq \operatorname{supp}(n)$, then $n' = \lambda n$ for some $\lambda \in \mathbb{R}$. Equivalently, n is an element of $\ker A$ with inclusion minimal support.

A *circuit* is the support set of a circuit vector. The minimality condition implies that the circuit determines its corresponding circuit vectors up to a multiple. A set C is called *circuit basis* if it contains precisely one circuit vector for every circuit. It is easy to see that a circuit basis of $\ker A$ spans $\ker A$. However, in general the circuit vectors are not linearly independent.

If we replace n by a nonzero multiple, then each equation of (1.22) is replaced by one that is equivalent over the nonnegative real numbers. Therefore all systems of equations corresponding to any circuit basis C are equivalent.

THEOREM 1.3.6. *Let $\mathcal{E}_A$ be an exponential family. Then $\overline{\mathcal{E}_A}$ equals the set of all probability distributions that satisfy*

$$p^{c^+} = p^{c^-} \text{ for all } c \in C, \tag{1.35}$$

where C is a circuit basis of A.

The proof is based on the following two lemmas, which are basic facts in oriented matroid theory:

LEMMA 1.3.7. *For every vector $n \in \ker A$ there exists a sign-consistent circuit vector $c \in \ker A$, i.e. if $c(i) \neq 0$ then $\operatorname{sgn} c(i) = \operatorname{sgn} n(i)$ for all $i \in [m]$.*

PROOF. Let c be a vector with inclusion-minimal support that is sign-consistent with n and satisfies $\operatorname{supp}(c) \subseteq \operatorname{supp}(n)$. If c is not a circuit, then there exists a circuit c' with $\operatorname{supp}(c') \subseteq \operatorname{supp}(c)$. Using a suitable linear combination $c + \alpha c'$, $\alpha \in \mathbb{R}$, we can obtain a contradiction to the minimality of c. □

LEMMA 1.3.8. *Every vector $n \in \ker A$ is a finite sign-consistent sum of circuit vectors $n = \sum_{i=1}^r c_i$, i.e. if $c_i(j) \neq 0$ then $\operatorname{sgn} c_i(j) = \operatorname{sgn} n(j)$ for all $j \in [m]$.*

PROOF. Use induction on the size of $\operatorname{supp}(n)$. In the induction step, use a sign-consistent circuit, as in the last lemma, to reduce the support. □

PROOF OF THEOREM 1.3.6. By Theorem 1.3.1 it suffices to show: If $p \in \mathbb{R}^m$ satisfies (1.35), then it also satisfies $p^{n^+} = p^{n^-}$ for all $n \in \ker A$. Write $n = \sum_{i=1}^r c_i$ as a sign-consistent sum of circuits c_i, as in the last lemma. Without loss of generality we assume $c_i \in C$ for all i. Then $n^+ = \sum_{i=1}^r c_i^+$ and $n^- = \sum_{i=1}^r c_i^-$. Hence p satisfies

$$p^{n^+} - p^{n^-} = p^{\sum_{i=2}^r c_i^+} \left(p^{c_1^+} - p^{c_1^-}\right) + \left(p^{\sum_{i=2}^r c_i^+} - p^{\sum_{i=2}^r c_i^-}\right) p^{c_1^-}, \tag{1.36}$$

where the first summand is zero by assumption. Now, the theorem follows by a finite induction on the number of circuits in the decomposition of n. □

The key in the proof of the theorem is that over the nonnegative real numbers we can take roots on both sides of an equation. This is the fundamental difference between algebraic geometry, and our statistical considerations.

REMARK 1.3.9 (Implicit characterization of the open family $\mathcal{E}_A$). It is easy to see that for the open family $\mathcal{E}_A$ the equations corresponding to a basis of the kernel of A suffice in the following sense: Let Z_1 be the solutions to the equations (1.35) and Z_2 be the solutions to $p^{b^+} - p^{b^-}$, where the exponents b run through a basis of $\ker A$. Then we have equality of their intersections with the open probability simplex: $Z_1 \cap \mathcal{P}_m = Z_2 \cap \mathcal{P}_m$. Namely, if p is strictly positive, we can take the component-wise logarithm $\log(p)$ and (1.22) holds if and only if

$$\sum_{i\in[m]} n^+(i) \log p(i) - \sum_{i\in[m]} n^-(i) \log p(i) = 0, \qquad \text{for all } n \in \ker A. \tag{1.37}$$

This already holds if $\sum_i b(i) \log p(i) = 0$ for all vectors b in a basis of $\ker A$.

EXAMPLE 1.3.10. This example was provided by Johannes Rauh and demonstrates the necessity of a circuit basis in the statement of Theorem 1.3.6. Let $1 \neq \alpha > 0$ and consider

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ -\alpha & 1 & 0 & 0 \end{pmatrix}, \tag{1.38}$$

The kernel is then spanned by

$$v_1 = (1, \alpha, -1, -\alpha)^T \text{ and } v_2 = (1, \alpha, -\alpha, -1)^T. \tag{1.39}$$

These two generators correspond to the two relations

$$p(1)p(2)^\alpha = p(3)p(4)^\alpha, \text{ and } p(1)p(2)^\alpha = p(3)^\alpha p(4). \tag{1.40}$$

It follows immediately that

$$p(3)p(4)^\alpha = p(3)^\alpha p(4). \tag{1.41}$$

If $p(3)p(4)$ is not zero, then we conclude $p(3) = p(4)$. However, on the boundary this does not follow from equations (1.40): Possible solutions to these equations are given by

$$p_a = (0, a, 0, 1-a) \text{ for } 0 \leq a < 1. \tag{1.42}$$

However, p_a does not lie in the closure of the exponential family $\overline{\mathcal{E}_A}$, since all members of $\mathcal{E}_A$ satisfy $p(3) = p(4)$. A circuit basis of A is given by the following vectors:

$$(0,0,1,-1)^T \qquad p(3) = p(4), \tag{1.43a}$$

$$(1,\alpha,0,-1-\alpha)^T \qquad p(1)p(2)^\alpha = p(4)^{1+\alpha}, \tag{1.43b}$$

$$(1,\alpha,-1-\alpha,0)^T \qquad p(1)p(2)^\alpha = p(3)^{1+\alpha}. \tag{1.43c}$$

REMARK 1.3.11. Using arguments from matroid theory the number of circuits can be shown to be less or equal than $\binom{m}{r+2}$, where r is the dimension of the exponential family $\mathcal{E}_A$, see [DSL04]. This gives an upper bound on the number of implicit equations necessary to describe $\overline{\mathcal{E}_A}$. Note that $\binom{m}{r+2}$ is usually much larger than the codimension $m - r - 1$ of $\mathcal{E}_A$ in the probability simplex. In contrast to this, if we only want to find an implicit description of all probability distributions of $\mathcal{E}_A$, which have full support, then $m - r - 1$ equations are enough: We can test $p \in \mathcal{E}_A$ by checking whether $\log(p)$ lies in the column span of A as in Remark 1.3.9.

It turns out that even in the boundary the number of equations can be reduced further. In general we do not need all circuits for the implicit description of $\overline{\mathcal{E}_A}$. For instance, in Example 1.3.10, the equations (1.43b) and (1.43c) are equivalent given (1.43a), i.e. we only need two of the three circuits to describe $\overline{\mathcal{E}_A}$. Unfortunately it is complicated to find a minimal subset of circuits that characterizes the closure of the exponential family. In the algebraic case discussed in Section 1.4 this question is equivalent to determining a minimal generating set of the circuit ideal among the circuits.

We give a simple algorithm that computes a circuit basis of A:

ALGORITHM 1.3.12 ([Stu02], Chapter 8). *Let r denote the rank of A. Without loss of generality we assume that A is an $r \times m$ matrix, i.e. redundant rows have been removed. For any $(r+1)$-subset $\tau = \{\tau_1, \dots, \tau_{r+1}\} \subseteq [m]$ of columns of A compute the vector:*

$$c_\tau := \sum_{i=1}^{r+1} (-1)^i \det(A_{\tau \setminus \{\tau_i\}}) e_{\tau_i}, \tag{1.44}$$

where e_{τ_i} is the canonical unit vector and $A_\sigma = (A_i)_{i \in \sigma}$ is the submatrix of columns with indices in $\sigma \subseteq [m]$. Any nonzero c_τ is a circuit vector of A and any circuit vector of A is proportional to a vector c_τ.

PROOF. We first prove that $Ac_\tau = A_\tau c_\tau = 0$, i.e. c_τ is orthogonal to any row of A_τ. Consider A', the $(r+1)\times(r+1)$ matrix which is received by doubling a row under consideration in A_τ. Naturally $\det(A') = 0$ and Laplace expansion of A' in the added row gives that $A_\tau c_\tau = 0$. Minimality of the support of c_τ is built into the formula (1.44). Namely each A_τ either has a unique circuit, in which case c_τ is proportional to all circuit vectors that it supports, or $\operatorname{rk} A_\tau < r$ and c_τ is zero. It remains to see that every circuit vector of A occurs among the c_τ. No circuit vector can have support larger than $r+1$ and all the $r+1$ subsets are considered in the algorithm. For each circuit σ, A_σ can be extended to an $r\times(r+1)$ matrix of rank r with columns τ. Then c_τ recovers the circuit σ. $\square$

Next we have another look at the relation of the implicit description and the possible support sets. Not only do the equations in Theorem 1.3.6 allow us to test a given p, together with Proposition 1.3.5 we can also test a possible support by testing the uniform distribution supported on it. Using the oriented matroid, that is the circuits together with an orientation we can further simplify this characterization. Let $S \subseteq \mathcal{X}$ be any set of configurations. Is there a probability distribution $p \in \overline{\mathcal{E}_A}$ satisfying $\operatorname{supp}(p) = S$? Answering this question is characterizing the set

$$\mathcal{S}(A) := \{\operatorname{supp}(p) : p \in \overline{\mathcal{E}_A}\} \subseteq 2^{[m]}, \tag{1.45}$$

and thus answering the support set problem from the introduction. The previous considerations give the following characterization: A nonempty set $S \subseteq [m]$ is the support set of some distribution $p \in \overline{\mathcal{E}_A}$ if and only if the following holds for all circuit vectors $n \in \ker A$:

- $\operatorname{supp}(n^+) \subseteq S$ if and only if $\operatorname{supp}(n^-) \subseteq S$.

Obviously, this condition does not depend on the circuits themselves, but on the supports of their positive and negative part. This is the combinatorial essence of an *oriented matroid*. In order to formalize this, consider the map

$$\operatorname{sgn}\colon n \mapsto (\operatorname{supp}(n^+), \operatorname{supp}(n^-)), \tag{1.46}$$

which associates to each vector a pair of disjoint subsets of $[m]$. Such a pair of disjoint subsets shall be called a *signed subset* of $[m]$ in the following. Alternatively, signed subsets (B^+, B^-) can also be represented as sign vectors $X \in \{-1, 0, +1\}^{[m]}$, where

$$X(i) = \begin{cases} +1, & \text{if } i \in B^+, \\ -1, & \text{if } i \in B^-, \\ 0, & \text{otherwise.} \end{cases} \tag{1.47}$$

In this representation, sgn corresponds to the usual sign mapping extended to vectors. As a slight abuse of notation, we do not make a difference between these two representations in the following.

The signed subset $\operatorname{sgn}(c)$ corresponding to a circuit vector $c \in \ker A$ shall be called an *oriented circuit.* The set of all oriented circuits is denoted by

$$\mathcal{C}(A) := \pm\operatorname{sgn}(C) = \{\operatorname{sgn}(c) : c \in C \text{ or } c \in -C\}, \tag{1.48}$$

where C is a circuit basis of A. We immediately have the following

THEOREM 1.3.13. *Let S be a nonempty subset of $[m]$. Then $S \in \mathcal{S}$ if and only if the following holds for all signed circuits $(B^+, B^-) \in \mathcal{C}(A)$:*

$$B^+ \subseteq S \quad \Leftrightarrow \quad B^- \subseteq S. \tag{1.49}$$

COROLLARY. *If two matrices A_1, A_2 satisfy $\mathcal{C}(A_1) = \mathcal{C}(A_2)$ then the possible support sets of the corresponding exponential families $\mathcal{E}_{A_1}$ and $\mathcal{E}_{A_2}$ coincide.*

According to Remark 1.3.11, Theorem 1.3.13 gives us up to $\binom{m}{r+2}$ conditions on the support. Usually, some of these conditions are redundant, but it is not easy to see a priori, which conditions are essential. Of course, a necessary condition for a subset S of $\mathcal{X}$ to be a support set of a distribution contained in $\overline{\mathcal{E}_A}$ is condition (1.49) restricted to pairs from a subset $\mathcal{H} \subseteq \mathcal{C}(A)$. For example, one can take $\mathcal{H} = \operatorname{sgn}(B)$, where B is a finite subset of $\ker A$, such as a basis.

EXAMPLE 1.3.14. Let us continue Example 1.3.10. From the circuits we deduce the following implications:

$$p(3) \neq 0 \iff p(4) \neq 0, \tag{1.50a}$$

$$p(1) \neq 0 \text{ and } p(2) \neq 0 \iff p(4) \neq 0, \tag{1.50b}$$

$$p(1) \neq 0 \text{ and } p(2) \neq 0 \iff p(3) \neq 0. \tag{1.50c}$$

Again, as above, the last two implications are equivalent given the first. From this it follows that the possible support sets in this example are $\{1\}$, $\{2\}$ and $\{1, 2, 3, 4\}$. From the spanning set (1.39) we only obtain the implication

$$p(1) \neq 0 \text{ and } p(2) \neq 0 \iff p(3) \neq 0 \text{ and } p(4) \neq 0. \tag{1.51}$$

We conclude this section with an example where a complete characterization of the face lattice of the convex support, and thus of the possible supports, is easily achievable. The convex support will be the well known *cyclic polytope*. Define the *moment curve* in $\mathbb{R}^d$ by

$$\boldsymbol{x} : \mathbb{R} \to \mathbb{R}^d, \qquad t \mapsto \boldsymbol{x}(t) := \left(t, t^2, \cdots, t^d\right)^T. \tag{1.52}$$

The *d-dimensional cyclic polytope with n vertices* is

$$C(d, n) := \operatorname{conv} \{\boldsymbol{x}(t_1), \ldots, \boldsymbol{x}(t_n)\}, \tag{1.53}$$

the convex hull of $n > d$ distinct points ($t_1 < t_2 < \ldots < t_n$) on the moment curve. The face lattice of a cyclic polytope can be described using *Gale's evenness condition*:

THEOREM 1.3.15. *A d-subset F of the vertices of $C(d, n)$ forms a facet if and only if for any $i, j \notin F$ with $i < j$, the number of $k \in F$ with $i < k < j$ is even.*

A nice proof using the Vandermonde determinant is found in [Zie94]. The cyclic polytope is of high importance because of its extremal properties. It is simplicial, all its proper faces are simplices, and neighborly, i.e. the convex hull of any $\lfloor \frac{d}{2} \rfloor$ vertices is a face of $C(n, d)$, but even better, one has

THEOREM 1.3.16 (Upper Bound Theorem). *If P is a d-dimensional polytope with $n = f_0$ vertices, then for every k it has at most as many k-dimensional faces as the cyclic polytope $C(d, n)$:*

$$f_k(P) \leq f_k(C(d, n)), \quad k = 0, \ldots, d. \tag{1.54}$$

If equality holds for some k with $\lfloor \frac{d}{2} \rfloor \leq k \leq d$ then P is neighborly.

Theorem 1.3.16 was conjectured by Motzkin in 1957 and its proof has a long and complicated history. The final result for polytopes is due to McMullen [McM70]. Later, in 1975, Richard Stanley showed that the upper bound theorem also holds for the f-vectors of triangulations of the sphere. This result was derived by constructing a certain ring for each simplicial complex, such that the upper bound holds whenever this ring, now called the Stanley-Reisner ring, has the Cohen-Macaulay property. These results mark the beginning of the field of combinatorial commutative algebra [Sta96; MS05].

The Upper Bound Theorem shows that the exponential families "modeled after cyclic polytopes", have the largest number of support sets among all exponential families with the same dimension and the same number of vertices.

Finally, we consider a cyclic polytope of dimension two that answers the question for the exponential family of smallest dimension and containing all the vertices of the probability simplex. The following construction using the moment curve appeared for instance in [MA03].

EXAMPLE 1.3.17. Consider the matrix A, whose columns are the points on the two-dimensional moment curve, augmented with row $(1, \dots, 1)$:

$$A := \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & 2 & 3 & \dots & m \\ 1 & 4 & 9 & \dots & m^2 \end{pmatrix}. \tag{1.55}$$

This matrix defines a two-dimensional exponential family. To approximate an arbitrary extreme point δ_j of the probability simplex, consider the parameter vector $\theta = (j^2, -2j, 1)^T$, giving rise to probability measures $p_{\beta,\theta} = \frac{1}{Z}\exp(-\beta\theta^T A)$. Since $\theta^T A_i = (i-j)^2$, we get that $\lim_{\beta\to\infty} p_{\beta,\theta} = \delta_j$. By definition, $\mathrm{cs}(\overline{\mathcal{E}_A})$ is the cyclic polytope $C(2,m)$ and its face lattice can be derived from Gale's evenness condition.

Summarizing we see that cyclic polytopes, owing to their extremal properties, have something to offer not only for convex geometry, but also for statistics. These examples conclude our discussion of the implicit representation of general exponential families. The relation to oriented matroid theory here might be only visible to the experienced eye. We will sketch the theory in the following section and also refer the reader [BVS+93] for the full theory.

1.3.1. Realizable Oriented Matroids. A realizable oriented matroid is the combinatorial data of a vector configuration. As exponential families are defined by point configurations, we consider only realizable oriented matroids. The content of this section are basics of oriented matroid theory, which also appeared in [RKA09]. The proofs are omitted or only sketched. Statements that are necessary for our application will be proved directly in the remaining parts.

Let E be a finite set and $\mathcal{C}$ a nonempty collection of signed subsets of E. For every signed set $X = (X^+, X^-)$ of E we let $\underline{X} := X^+ \cup X^-$ denote the *support* of X. Furthermore, the *opposite signed set* is $-X = (X^-, X^+)$. Then the pair $(E, \mathcal{C})$ is called an *oriented matroid* if the following conditions are satisfied:

- **(C1)** $\mathcal{C} = -\mathcal{C}$, (*symmetry*)
- **(C2)** for all $X, Y \in \mathcal{C}$, if $\underline{X} \subseteq \underline{Y}$, then $X = Y$ or $X = -Y$, (*incomparability*)
- **(C3)** for all $X, Y \in \mathcal{C}$, $X \neq -Y$, and $e \in X^+ \cap Y^-$ there is a $Z \in \mathcal{C}$ such that $Z^+ \subseteq (X^+ \cup Y^+) \setminus \{e\}$ and $Z^- \subseteq (X^- \cup Y^-) \setminus \{e\}$. (*weak elimination*)

In this case each element of $\mathcal{C}$ is called a *signed circuit*.

Note that to every oriented matroid $(E, \mathcal{C})$ we have an associated unoriented matroid (E, C), called the *underlying matroid*, where

$$C = \{X^+ \cup X^- = \mathrm{supp}(X) : X \in \mathcal{C}\} \tag{1.56}$$

is the set of *circuits* of (E, C). In this way oriented matroids can be considered as ordinary matroids endowed with an additional structure, namely a *circuit orientation*, which assigns two opposite signed circuits $\pm X \in \mathcal{C}$ to every circuit $X \in C$.

The most important example of an oriented matroid is that of a matrix $A \subseteq \mathbb{R}^{d\times m}$. In this case let $E = [m]$, and

$$\mathcal{C} = \big\{(\operatorname{supp}(n^+), \operatorname{supp}(n^-) : n \in \ker A \text{ has inclusion minimal support}\big\}. \tag{1.57}$$

This example is so important that oriented matroids that arise in this way are given a name: An oriented matroid is called *realizable* if it is induced by some matrix A. Note that this definition, like the definition of an exponential family in (1.7) depends only on the rowspace of A. In this context, which is dominant in the following, we also call the vectors $n \in \ker A$ *circuit vectors*, or, slightly abusing notation, simply *circuits* if their support is minimal among the supports of elements in $\ker A$.

The only axiom that is not trivially fulfilled for this example is **(C3)**. However, if we drop the minimality condition and let $\mathcal{V} = \{(\operatorname{supp}(n^+), \operatorname{supp}(n^-) : n \in \ker A\}$, then it is easily seen that $\mathcal{V}$ satisfies **(C3)**. Thus $(E, \mathcal{C})$ satisfies **(C3)** by the following proposition:

Proposition 1.3.18. *Let $\mathcal{V}$ be a nonempty collection of signed subsets of E satisfying* **(C1)** *and* **(C3)**. *Write* $\operatorname{Min}(\mathcal{V})$ *for the minimal elements of $\mathcal{V}$ (with respect to inclusion of supports).*

(1) For any $X \in \mathcal{V}$ there is $Y \in \operatorname{Min}(\mathcal{V})$ such that $Y^+ \subseteq X^+$ and $Y^- \subseteq X^-$.
(2) $\operatorname{Min}(\mathcal{V})$ *is the set of circuits of an oriented matroid.*

Proof. [BVS+93], Proposition 3.2.4. □

This illustrates how **(C2)** corresponds to the minimality condition. It is possible to define oriented matroids without this minimality condition using the following construction:

For two signed subsets X, Y of E define the *composition* $X \circ Y$ of X and Y as

$$(X \circ Y)^+ := X^+ \cup (Y^+ \setminus X^-), \qquad (X \circ Y)^- := X^- \cup (Y^- \setminus X^+). \tag{1.58}$$

Note that this operation is associative but not commutative in general. A composition $X \circ Y$ is *conformal* if X and Y are *sign-consistent*, i.e. $X^+ \cap Y^- = \emptyset = X^- \cap Y^+$, and $X \circ Y = Y \circ X$.

An *o.m. vector* of an oriented matroid is any composition of an arbitrary number of circuits. In [BVS+93], o.m. vectors are simply called vectors. The name "o.m. vector" has been proposed by F. Matúš to avoid confusion. The set of o.m. vectors shall be denoted by $\mathcal{V}$. If the oriented matroid comes from a matrix A, then $\mathcal{V}$ equals the set $\mathcal{V}$ from above.

The above proposition implies that an oriented matroid can be defined as a pair $(E, \mathcal{V})$, where $\mathcal{V}$ is a collection of signed subsets satisfying **(C1)**, **(C3)** and

(V0) $\emptyset \in \mathcal{V}$,
(V2) for all $X, Y \in \mathcal{V}$ we have $X \circ Y \in \mathcal{V}$,

Note that in the realizable case linear combinations of vectors correspond to composition of their sign vectors in the following sense:

$$\operatorname{sgn}(n + \epsilon n') = \operatorname{sgn}(n) \circ \operatorname{sgn}(n'), \quad \text{for } \epsilon > 0 \text{ small enough.} \tag{1.59}$$

Now Lemmas 1.3.7 and 1.3.8 correspond to the following two lemmas

Lemma 1.3.7'. *For every o.m. vector Y there exists a sign-consistent signed circuit X such that $\underline{X} \subseteq \underline{Y}$.*

Lemma 1.3.8'. *Any o.m. vector is a conformal composition of circuits.*

To every matrix A we associate its convex support. Many properties of this polytope can be translated into the language of oriented matroids. This yields constructions which also make sense if the oriented matroid is not realizable. In order to make this more precise, we need the notion of the dual oriented matroid, to be discussed only in the realizable case here.

Assume that the matrix A has the constant vector $(1, \dots, 1)$ in its rowspace. Then all the column vectors a_x lie in a hyperplane $l_1 = 1$, for some dual vector $l_1 \in (\mathbb{R}^d)^*$. In the general case, this can always be achieved by adding another dimension. Technically we require that the face lattice of the polytope spanned by the columns of A is combinatorially equivalent to the face lattice of the cone over the columns. See also the remarks before Definition 1.3.3.

For every dual vector $l \in (\mathbb{R}^d)^*$ let $N_l^+ := \{i \in [m] : l(a_i) > 0\}$ and $N_l^- := \{i \in [m] : l(a_i) < 0\}$. This way we can associate a signed subset $\operatorname{sgn}^*(l) := (N_l^+, N_l^-)$ with l. The signed subset $\operatorname{sgn}^*(l)$ is called a *covector*. Let $\mathcal{L}$ be the set of all covectors. If the signed subset (N_l^+, N_l^-) has minimal support (i.e. "many" vectors a_i lie on the hyperplane $l = 0$), then l is called a *cocircuit vector*, and $\operatorname{sgn}^*(l)$ is called a *signed cocircuit*. The collection of all signed cocircuits shall be denoted by $\mathcal{C}^*$.

Lemma 1.3.19. *Let $(E, \mathcal{C})$ be an oriented matroid induced by a matrix A. Then $(E, \mathcal{C}^*)$ is an oriented matroid, called the* dual oriented matroid*.*

Proof. See Section 3.4 of [BVS+93]. □

Note that the faces of the polytope correspond to hyperplanes such that all vertices lie on one side of this hyperplane, compare Definition 1.3.3. Thus the faces of the polytope are in a one-to-one relation with the positive covectors, i.e. the covectors $X = (X^+, X^-)$ such that $X^- = \emptyset$. The face lattice of the polytope can be reconstructed by partially ordering the positive covectors by inclusion of their supports; however, the relation needs to be inverted: Covectors with small support correspond to faces that contain many vertices. The empty face corresponds to the covector $T := ([m], \emptyset)$ and can be given by the dual vector l_1, defining the hyperplane containing all a_x.

The facts just discussed apply to all abstract oriented matroids such that $T = ([m], \emptyset)$ is a covector. Such an oriented matroid is usually called *acyclic*. Thus a face of an acyclic oriented matroid is any positive covector. A *vertex* is a maximal positive covector X in $\mathcal{L} \setminus \{T\}$, i.e. if $\underline{X} \subseteq \underline{Y}$ for some positive covector $Y \in \mathcal{L} \setminus \{X\}$, then $Y = T$.

In this setting we have the following result, which corresponds to the second statement of Lemma 1.3.4:

Proposition 1.3.20 (Las Vergnas). *Let $(E, \mathcal{C})$ be an acyclic oriented matroid. For any subset $F \subseteq E$ the following are equivalent:*

- *F is a face of the oriented matroid.*
- *For every signed circuit $X \in \mathcal{C}$, if $X^+ \subseteq F$ then $X^- \subseteq F$.*

Proof. The proof of Proposition 9.1.2 in [BVS+93] applies. Note that the statement of Proposition 9.1.2 includes an additional assumption which is never used in the proof. □

By means of the moment map, this proposition can be used to derive Theorem 1.3.13: Every face of the convex support corresponds to a possible support set of an exponential family, and the proposition links this to the signed circuits of the corresponding oriented matroid. Finally, the corollary to Theorem 1.3.13 can be rewritten as

Corollary. *The possible support sets of two exponential families coincide if they have the same oriented matroids.*

Unfortunately, this correspondence is not one-to-one: Different oriented matroids can yield the same face lattice, i.e. combinatorially equivalent polytopes. A simple example is given by a regular and a nonregular octahedron as described in [Zie94]. The special case in which this does not happen has a name: an oriented matroid is called *rigid* if its positive covectors (i.e. its face lattice) determine all covectors (i.e. the whole oriented matroid). Still, the corollary implies that the instruments of the theory of oriented matroids suffice to describe the support sets of an exponential family.

Remark 1.3.21 (Duality). There are two main reasons why the theory of oriented matroids (as well as the theory of ordinary matroids) is considered important. First, it yields an abstract framework which allows to describe a multitude of different combinatorial questions in a unified manner. This, of course, does not in itself lead to any new theorem. The second reason is that the theory provides the important tool of matroid duality. It turns out that the dual of a realizable matroid is again realizable: If A is a matrix representing an oriented matroid $(E, \mathcal{C})$, then any matrix A^* such that the rows of A^* span the orthogonal complement of the row span of A represents the oriented matroid $(E, \mathcal{C}^*)$. To motivate the importance of this construction we sketch its implications for the case that the oriented matroid comes from a polytope. In this case the duality is known under the name *Gale transform* [Zie94, Chapter 6]. A d-dimensional polytope with N vertices can be represented by N vectors in $\mathbb{R}^{d+1}$ lying in a hyperplane. These vectors form a $(d+1) \times N$-matrix A. Now we find an $(N-d-1) \times N$-matrix A^* as above, so the dual matroid is represented by a configuration of N vectors in $\mathbb{R}^{N-d-1}$. This means that this construction allows us to obtain a low-dimensional image of a high-dimensional polytope, as long as the number of vertices is not much larger than the dimension. This method has been used for example in [Stu88] in order to construct polytopes with quite unintuitive properties, leading to the rejection of some conjectures. Furthermore, oriented matroid duality makes it possible to classify polytopes with “few vertices” by classifying vector configurations.

The notion of dimension generalizes to arbitrary oriented matroids (and ordinary matroids). In the general setting one usually talks about the *rank* of a matroid, which is defined as the maximal cardinality of a subset $E \subseteq F$ such that E contains no support of a signed circuit. In this sense duality exchanges examples of high rank and low rank, where “high” and “low” is relative to $|E|$.

Remark 1.3.22 (Exponential Families from Vector Configurations). Having presented the general theory of vector configurations it is natural to view exponential families in this setting. Consider the oriented matroid associated to the vector configuration $\{v_1, \ldots, v_m\} \subseteq \mathbb{R}^{d+1}$ by letting $A \in \mathbb{R}^{(d+1)\times m}$ have columns $v_i, i \in [m]$, and defining $\mathcal{C}(A)$ as in (1.57). An exponential family $\overline{\mathcal{E}_A}$, associated to the vector configuration, is instantly defined. It is contained in the probability simplex $\overline{\mathcal{P}_m}$. This view on exponential families will hopefully lead to a better understanding of the underlying combinatorics. A discrete exponential family (with uniform distribution as its base measure) is characterized by a point configuration. Its combinatorics depends only on the information in the point configuration: its oriented matroid. This is the spirit in which [RKA09] was written.

1.4. Integer Valued Sufficient Statistics and Toric Varieties

In this section we will connect the general theory, as developed above, to the case that is considered in algebraic statistics. There the matrix A is assumed to take only rational values, and since the constant vector is assumed to lie in the row space, we can restrict to the case where it has only nonnegative integer entries. This very mild assumption allows us to enter the world of toric geometry. A classical text on the subject is [Ful93] and a more modern reference will be the upcoming book [CLS09]. Proofs of the statements will be omitted.

1.4.1. Toric Ideals and Varieties. The basic idea of toric geometry is that a certain geometric object, e.g. a manifold or variety, is modeled as a fiber bundle in which the fibers are multi-dimensional tori. Here we focus on toric varieties whose base spaces are convex polytopes and whose tori are products of the *algebraic torus* $\Bbbk^* := \Bbbk \setminus \{0\}$, where $\Bbbk$ is a field. For statistics we are interested in the real nonnegative part of a toric variety, which is essentially only the polytope called the convex support above. A reference is [Stu96].

In the following let $\Bbbk$ be a field of characteristic zero. Typical examples are the rational numbers $\mathbb{Q}$ or the complex numbers $\mathbb{C}$. Consider the polynomial ring $\Bbbk[\boldsymbol{x}] := \Bbbk[x_1, \ldots, x_m]$ in variables $x_1, \ldots, x_m$. Fix an *integer valued* matrix $A \in \mathbb{Z}^{d\times m}$ which has the constant vector $(1, 1, \ldots, 1)$ in its rowspace. For general affine toric varieties, this assumption could be dropped. Here we are only interested in the case of *projective* toric varieties that arise from matrices whose columns lie in a common hyperplane. Again, denote $\{a_i : i \in [m]\}$ the columns of A. Consider additionally the Laurent polynomial ring in d invertible variables: $\Bbbk[\boldsymbol{t}^{\pm 1}] := \Bbbk[t_1, \ldots, t_d, t_1^{-1}, \ldots, t_d^{-1}]$. The matrix A defines a homomorphism of $\Bbbk$-algebras:

$$\phi_A : \Bbbk[\boldsymbol{x}] \to \Bbbk[\boldsymbol{t}^{\pm 1}], \quad x_i \mapsto \boldsymbol{t}^{a_i} := \prod_{j=1}^{d} t_j^{a_{ji}}. \tag{1.60}$$

Definition 1.4.1. The ideal $I_A := \ker \phi_A$ is called the *toric ideal* of the matrix A. Its variety $V(I_A)$ is called a *toric variety*.

With this definition I_A is a homogeneous prime ideal, hence the projective variety $V(I_A)$ is irreducible. It equals the Zariski closure of $\left\{(\boldsymbol{t}^{a_1}, \ldots, \boldsymbol{t}^{a_m}) : \boldsymbol{t} \in (\Bbbk^*)^d\right\} \subseteq \Bbbk^m$. Written this way, we see the monomial parameterization, which marks the connection point to the theory of exponential families. Consider the exponential family as defined in (1.7). For a fixed p, corresponding to a parameter vector $c \in \mathbb{R}^d$, an elementary probability takes the form

$$p(j) = \frac{1}{Z_c} \exp\left(\sum_{i=1}^{d} c_i a_{ij}\right) = \prod_{i=1}^{d} t_i^{a_{ij}}, \tag{1.61}$$

where $t_i = \exp(c_i)$ and the normalization is "hidden" in the parameter for the constant row. A countably infinite generating set of the toric ideal I_A is given by

$$I_A = \langle x^u - x^v : u, v \in \mathbb{N}^m, Au = Av\rangle\,, \tag{1.62}$$

where the notation $\langle . \rangle$ indicates the ideal generated by the given equations. The ideal I_A should be compared to the equations in the statement of Theorem 1.3.1. In fact, in [GMS06] it is shown that the uncountably many nonpolynomial equations in Theorem 1.3.1 can be reduced to the countably many in (1.62), provided that A takes only nonnegative integer values. The proof of equivalence of (1.62) and the definition is simple and can be found in Lemma 4.1 of [Stu96]. This book also contains a collection of results about generating sets and Gröbner bases of toric ideals.

Next we want to see the torus acting on the toric variety. Let $\Bbbk = \mathbb{C}$, the matrix A induces an action of the algebraic torus $(\mathbb{C}^*)^d$ on $\mathbb{C}^m$ by considering the a_i for weights:

$$\psi : (\mathbb{C}^*)^d \times \mathbb{C}^m \to \mathbb{C}^m, \quad (t, c) \mapsto (c_1 t^{a_1}, \ldots, c_m t^{a_m}). \tag{1.63}$$

Since the a_i are considered to lie in an affine hyperplane this gives an induced action of $(\mathbb{C}^*)^d$ on projective $(m-1)$-space $\mathbb{P}^{m-1}$. The *projective* toric variety defined above is the closure of the orbit of the point $(1, 1, \ldots, 1)$. This can be taken as a definition too, then one has to prove that the vanishing ideal of this orbit is I_A. The torus action should be compared to a natural affine action on the probability simplex [MR93]. Let $\mathbb{R}\mathbb{1}$ denote the set of constant functions in $\mathbb{R}^m$. Because of the normalization of probability vectors, $\mathbb{R}^m/\mathbb{R}\mathbb{1}$ bijectively parametrizes $\mathcal{P}_m$. It acts on $\mathcal{P}_m$ via

$$\Psi : \mathcal{P}_m \times \mathbb{R}^m/\mathbb{1}\mathbb{R} \to \mathcal{P}_m, \quad (p, f + \mathbb{R}\mathbb{1}) \mapsto \frac{p\mathrm{e}^f}{\sum_{i=1}^m p(i)\mathrm{e}^{f(i)}}, \tag{1.64}$$

which is clearly well defined. With this definition, $\mathcal{P}_m$ becomes an affine space, so we could define exponential families simply as its affine subspaces. If $\mathcal{L} \subseteq \mathbb{R}^m$ is a linear subspace containing $\mathbb{R}\mathbb{1}$ then (1.5) takes the form

$$\mathcal{E}_{\mathcal{L}} = \{\Psi(\mathbb{1}, f + \mathbb{R}) : f \in \mathcal{L}\}, \tag{1.65}$$

which parallels the definition of a toric variety given below (1.63).

A final remark on the general theory is necessary. Care has to be taken when comparing these definitions with the classical ones in algebraic geometry, e.g. in [Ful93]. Those read as: *A toric variety is a* normal *variety X that contains an algebraic torus $T \cong (\Bbbk^*)^d$ as a dense open subset, together with an action $T \times X \to X$ that extends the natural action of T on itself.* In this definition normality is required, while the above definition allows toric varieties that are not normal. Chapter 13 of [Stu96] explains the differences and shows how to get the vector configuration of an embedded normal toric variety.

1.4.2. Integer Valued Sufficient Statistics. If A has only integer entries then every circuit vector is proportional to one with integer components (see Algorithm 1.3.12). In this case the corresponding equations (1.22) are polynomial, and the theorem implies that $\overline{\mathcal{E}_A}$ is the nonnegative real part of a *projective variety*, i.e. the solution set of homogeneous polynomials. It was already noted early that statistical models may possess polynomial functions vanishing on them. These are sometimes called invariants of the model. An easy example is the set of distributions of two independent binary random variables:

$$\{P(i,j) = P_1(i)P_2(j) : P_1, P_2 \text{ are univariate distributions}\}. \tag{1.66}$$

Naturally the polynomial equation $P(0,0)P(1,1) = P(1,0)P(0,1)$ holds. It is a polynomial function vanishing on the model, an invariant. In this simple case, the single equation characterizes all independent distributions: Every equation vanishing on the model is a multiple of this equation; we have found the complete set of invariants. A far reaching generalization of the idea to specify certain dependencies between discrete random variables is that of a hierarchical model, to which we devote Chapter 2. There we try to find complete sets of invariants, called Markov bases.

Returning to our general discussion, if we want to use the tools of commutative algebra and algebraic geometry, then it turns out that circuits are not the right object to consider. This is due to the fact that the geometric object, the exponential family that we have described using the circuits is not closed in the Zariski topology. If we additionally consider all complex solutions of equations corresponding to circuits we get a set which in general strictly contains the Zariski closure of the real solutions. One basic source of disagreement is that proportional circuits only yield equivalent equations if we consider them over the nonnegative reals, but we may obtain a different solution set if we allow negative real solutions or complex solutions. If $\ker A$ is an integer lattice, then there are countably infinitely many equations characterizing the Zariski closure of $\overline{\mathcal{E}_A}$. They define the ideal

$$I_A := \langle p^u - p^v : u, v \in \mathbb{N}^m, Au = Av \rangle \subseteq \mathbb{Q}[p_1, \dots, p_m], \tag{1.67}$$

which is the same as (1.62). Here $\mathbb{Q}[p_1, \dots, p_m]$ is the ring of polynomials in elementary probabilities and with rational coefficients. As noted above, this ideal is prime and its variety has a monomial parameterization. According to Hilbert's Basis Theorem there are finitely many equations that suffice to generate this ideal. The exponent vectors of these finitely many generators are called a Markov basis and have interesting applications in statistics [DS98; DSS09]. By the discussion above the set of circuits need not include a Markov basis.

Finding a Markov basis is in general a nontrivial task. In small examples it can be carried out with a computer algebra system, using the basic fact that the toric ideal (1.67) equals the saturation of the lattice basis ideal of $\ker A$ [HM09]. In contrast, it is easy to compute the circuits of a matrix using Algorithm 1.3.12.

A minimal Markov basis is usually much smaller than a circuit basis, and thus it is easier to handle. A specialized open source software for these tasks is 4ti2 [4ti207], which can compute circuits as well as Markov bases and was used intensively by the author.

Example 1.4.2. Consider the matrix

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{pmatrix}. \tag{1.68}$$

An easy calculation with Algorithm 1.3.12 shows that the circuits of A are

$$\begin{array}{ll} (0, 1, -2, 1), & (1, -2, 1, 0) \\ (1, 0, -3, 2), & (2, -3, 0, 1). \end{array} \tag{1.69}$$

A computation with 4ti2 shows that $(1, -1, -1, 1)$ is an element of the minimal Markov basis, however, with its support of four elements it is not a circuit.

One can also look at the ideal generated by all polynomial equations induced by integer valued circuit vectors. This ideal is called the *circuit ideal*. By what was said above this ideal is in general smaller than the associated toric ideal, which contains the polynomial equations induced by all integer valued kernel vectors. Circuit ideals, i.e. binomial ideals whose generators' exponent vectors are the circuits of an integer matrix have been studied already in [ES96], whose results play a key role in Chapter 3. Further results illuminating the nice relations of circuit ideals and polyhedral geometry can be found in [BJT07].

Summarizing, in the algebraic setting, Theorem 1.3.1 remains valid if we replace "closure" by "Zariski closure" and $\ker A$ by the integer kernel $\ker_{\mathbb{Z}} A$. This fact was first noted in [DS98] and is one of the cornerstones of algebraic statistics.

1.4.3. Experimental Designs. In this section we briefly mention how the sufficient statistics matrices A are discussed within the framework of experimental design. A reference, and actually the first book on algebraic statistics, is [PRW01]. Assume we are given a fixed set of points $\mathcal{X} \subseteq \mathbb{R}^d$, the *design points*. These points correspond to values of parameters of some experimental setup where actual measurements can be made. Let $|\mathcal{X}| = m$, we wish to describe the space of real valued functions on $\mathcal{X}$. Commutative algebra offers the tools for this. Let $S = \mathbb{R}[x_1, \ldots, x_d]$ be a polynomial ring in d indeterminates and consider $I = I(\mathcal{X})$, the ideal of the zero-dimensional variety $\mathcal{X}$. This ideal can be computed by taking the intersection of m zero-dimensional ideals. The computer algebra system CoCoA[CoC] has an optimized function for this task. The quotient ring S/I is the real vector space of functions on $\mathcal{X}$. A basis consists of the set of standard monomials, which are defined as all monomials not contained in the initial ideal with respect to some term order. Now specifying a model can be done by selecting a finite subset $\{t_1, \ldots, t_r\}$ of this basis and considering the matrix A with columns indexed by $\mathcal{X}$ and rows indexed by $[r]$ and entries $A(x,i) = t_i(x)$. In this context the matrix A is called a *design matrix*. As an example consider $\mathcal{X} = \{\pm 1\}^3$. Its defining ideal is $\langle x^2-1, y^2-1, z^2-1\rangle \subseteq \mathbb{R}[x,y,z]$ and the given generators form a Gröbner basis. The standard monomials are $1, x, y, z, xy, xz, yz, xyz$. The design matrix is the Hadamard matrix

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{pmatrix}. \tag{1.70}$$

This design is called the *full-factorial design with 3 factors* and is only a very simple example. The theory of experimental designs can be understood in terms of commutative algebra, leading to new insights which are also subsumed under the name algebraic statistics [PRW01].

CHAPTER 2

Boundaries of Hierarchical Models

In this chapter we will specialize to a very important class of exponential families: hierarchical loglinear models. By imposing additional structure on the ground set we define an exponential family $\mathcal{E}_{\Delta,\boldsymbol{d}}$ to each choice of a simplicial complex Δ, and a vector of integers $\boldsymbol{d}$. Many important examples of discrete statistical models are hierarchical models; it is of great importance to derive properties of their boundaries. Hierarchical models are described by 0/1-matrices and thus we are in the setting of toric algebra. This observation is fundamental in algebraic statistics and has lead to many interesting theoretical insights and applications.

Let $A \in \mathbb{Z}^{d\times m}$ be a matrix with integer entries and denote $\ker_{\mathbb{Z}} A$ its integer kernel, a submodule of $\mathbb{Z}^m$. In statistics, more specifically for the computation of p-values in the analysis of contingency tables, one is interested in sampling from the set of integer valued tables with given fixed margins. An efficient procedure for this allows one to perform Fischer's exact test. One of the first applications of algebra in statistics is to construct connected Markov chains on the set of tables with fixed margins [DS98]. The set of elementary moves in such a Markov chain is called a *Markov basis*. Now, algebra comes into play through the observation that a finite set $M \subseteq \ker_{\mathbb{Z}} A$ forms a Markov basis if and only if its associated binomials generate the toric ideal from (1.67)

$$I_A = \left\langle p^{m^+} - p^{m^-} : m \in M \right\rangle . \tag{2.1}$$

For the above reasons, elements of $\ker A$ are called *moves*, and elements in a Markov basis are called *Markov moves*. We define the degree of a move as the sum of its positive components, which equals the sum of the negative components as $(1,\dots,1)$ is in the row space of A. The degree of the binomial $p^{m^+} - p^{m^-}$ then equals the degree of m.

In the first section we give the basic definitions and review some material from algebraic statistics. In Section 2.2 we show a construction of circuits of hierarchical models that give new insight into symmetry and support properties. These elementary circuits drive much of the following development. A relation to notions of independence is discussed and a lower bound on the supports of Markov moves is derived. Section 2.3 gathers material on Markov bases and also includes an introduction to MBDB, the Markov Bases Database. Section 2.4 is devoted to binary hierarchical models and their connection to coding theory. Finally the Chapter is concluded with the implications for the support set problem in Section 2.5.

2.1. Hierarchical Models

We are in the setting of Chapter 1, but in the following we consider a ground set $\mathcal{X}$ of compositional structure, i.e. $\mathcal{X} = \prod_{i=1}^{n} \mathcal{X}_i$, for some finite sets $\mathcal{X}_i$, $i = 1,\dots,n$. We motivate this with our main example: modeling dependencies in contingency tables. Consider a collection of n random variables taking values in nonempty, finite sets $\mathcal{X}_i, i = 1,\dots,n$. The identities of the elements of $\mathcal{X}_i$ play no further role. Therefore, defining $d_i := |\mathcal{X}_i|$, the information about these sets is completely specified by the sequence $\boldsymbol{d} := (d_1, \dots, d_n)$ of cardinalities. In the notation of Chapter 1 we have $m := \prod_{i=1}^{n} d_i$. The vector space of real valued functions is denoted $\mathbb{R}^{\mathcal{X}}$ and is, after choosing any order, isomorphic to $\mathbb{R}^m$. The open simplex of strictly positive probability distributions is denoted $\mathcal{P}(\mathcal{X})$, its closure $\overline{\mathcal{P}(\mathcal{X})}$. This is the set of joint distributions of the n random variables.

When using matrices to describe linear maps we assume the standard basis given by the functions

$$e^{(x)}(y) := \begin{cases} 1 & \text{if } y = x, \\ 0 & \text{otherwise,} \end{cases} \qquad x \in \mathcal{X}. \tag{2.2}$$

This allows us to consider for instance a matrix $A \in \mathbb{R}^{d \times m}$ as a linear map $\mathbb{R}^{\mathcal{X}} \to \mathbb{R}^d$. We denote $N := \{1, \dots, n\}$, and its power set as $2^N := \{B : B \subseteq N\}$. For a subset $B \subseteq N$ of the random variables, we denote its set of values as $\mathcal{X}_B := \prod_{i \in B} \mathcal{X}_i$. We have the natural projections

$$\begin{aligned} X_B : \mathcal{X} &\to \mathcal{X}_B \\ (x_i)_{i \in N} &\mapsto (x_i)_{i \in B} =: x_B. \end{aligned} \tag{2.3}$$

It is convenient to slightly abuse notation and denote x_B the projection of x to B, which depends on x, and by the same symbol an arbitrary element $x_B \in \mathcal{X}_B$. We frequently use *cylinder sets*, specified by a set $B \subseteq N$, and $y_B \in \mathcal{X}_B$, containing all elements that have y_B as their projection to B:

$$\{X_B = y_B\} := \{x \in \mathcal{X} : X_B(x) = y_B\}. \tag{2.4}$$

In statistics, a nonnegative integer valued vector $u \in \mathbb{N}_0^{\mathcal{X}}$ is called a *contingency table*. It can be thought of as a summary of a discrete sample from $\mathcal{X}$, where we record how often each $x \in \mathcal{X}$ occurred. We can further summarize such data by just looking at a subset $B \subseteq N$. To this end we define the marginal table $u_B \in \mathbb{N}_0^{\mathcal{X}_B}$ as the vector with components

$$u_B(x_B) := \sum_{y : X_B(y) = x_B} u(y), \qquad x_B \in \mathcal{X}_B. \tag{2.5}$$

We now define a hierarchical model by specifying interaction structures between the nodes $i \in N$. A convenient way to do so is by giving an abstract simplicial complex Δ on N [HS02; DS03]. A simplicial complex is closed under taking subsets, i.e. $A \in \Delta, B \subseteq A \Rightarrow B \in \Delta$. The elements of Δ elements are called *faces* and the *facets* $\mathcal{F}$ are defined as the inclusion maximal faces. By definition, Δ is completely determined by the list of facets. We use this fact and denote simplicial complexes using the *bracket notation* [Chr97], in which the facets are listed in brackets. For instance $\Delta = [12][13][23]$ is the bracket notation for

$$\Delta = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}\}. \tag{2.6}$$

The facets determine the *marginal map*:

$$\begin{aligned} A_\Delta : \mathbb{R}^{\mathcal{X}} &\to \bigoplus_{F \in \mathcal{F}} \mathbb{R}^{\mathcal{X}_F}, \\ u &\mapsto (u_F)_{F \in \mathcal{F}}. \end{aligned} \tag{2.7}$$

It is a linear map computing all marginal tables corresponding to facets. With respect to the canonical basis, the matrix representing this map, also denoted A_Δ, is the $k \times |\mathcal{X}|$

matrix

$$A_\Delta := (A_{(B,y_B),x})_{(B,y_B),x} \text{ where } A_{(B,y_B),x} := \begin{cases} 1 & \text{if } X_B(x) = y_B, \\ 0 & \text{otherwise.} \end{cases} \tag{2.8}$$

The rows of this matrix are indexed by pairs (B, y_B), where $B \in \mathcal{F}$ is a facet of Δ and $y_B \in \mathcal{X}_B$ is a configuration on B. Then k is defined as the number of such pairs. If relevant, we indicate the dependency on the vector $\boldsymbol{d}$ of cardinalities in $\mathcal{X}$ by writing $A_{\Delta,\boldsymbol{d}}$ or simply $A_{\boldsymbol{d}}$. Note that rows (B, y_B) corresponding to *faces* of Δ are linearly dependent on the columns corresponding to facets. In particular the row $(1, \ldots, 1)$, corresponding to the empty set, is contained in the row space of A_Δ. The columns of A_Δ are denoted $A_{\Delta,x}$ or A_x respectively. Now we are ready for the central definition:

Definition 2.1.1. The *open hierarchical model* for the simplicial complex Δ and cardinalities $\boldsymbol{d}$ is the exponential family of $A_{\Delta,\boldsymbol{d}}$:

$$\mathcal{E}_{\Delta,\boldsymbol{d}} := \mathcal{E}_{A_{\Delta.\boldsymbol{d}}} = \left\{ p \in \mathcal{P}(\mathcal{X}) : p(x) = Z_c^{-1} \exp\left(\langle c, A_{\Delta,x}\rangle\right) : c \in \mathbb{R}^k \right\}, \tag{2.9}$$

where the partition function Z_c is

$$Z_c := \sum_{x \in \mathcal{X}} \exp\left(\langle c, A_{\Delta,x}\rangle\right). \tag{2.10}$$

A *hierarchical model* is the closure of an open hierarchical model, denoted $\overline{\mathcal{E}_{\Delta,\boldsymbol{d}}}$. Whenever there is no ambiguity, we omit indices and use the symbols $\mathcal{E}, \mathcal{E}_\Delta, \mathcal{E}_{\boldsymbol{d}}, \overline{\mathcal{E}_\Delta}, \ldots$ The convex support $\operatorname{cs}(\overline{\mathcal{E}_{\Delta,\boldsymbol{d}}})$ of a hierarchical model is called its *marginal polytope*.

Remark 2.1.2. One can also define a hierarchical model through a factorization property

$$\mathcal{E}_{\Delta,\boldsymbol{d}} = \left\{ p \in \mathcal{P}(\mathcal{X}) : p(x) = \prod_{F \in \Delta} \phi_F(x_F) \right\}, \tag{2.11}$$

where the ϕ_F are functions depending on their arguments only through the set $F \subseteq N$. More formally $\phi_F(x_F, x_{N\setminus F}) = \phi_F(x_F, x'_{N\setminus F})$ for all $x_F \in \mathcal{X}_F$ and $x_{N\setminus F}, x'_{N\setminus F} \in \mathcal{X}_{N\setminus F}$. This definition is often used in the theory of graphical models [Lau96]. See also Example 2.1.5.

Example 2.1.3 (Two independent binary random variables). In the case of two binary random variables, we have $\mathcal{X} = \{(00), (01), (10), (11)\}$. Let $\Delta = [1][2]$, then the matrix A_Δ is given by

$$A_\Delta = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}. \tag{2.12}$$

The rows are ordered as $(\{1\}, 0), (\{1\}, 1), (\{2\}, 0), (\{2\}, 1)$. If Δ was the whole power set, A_Δ would be the (4×4)-identity matrix. The marginal polytopes are easily identified as a 2-dimensional square and a 3-dimensional simplex, respectively.

Example 2.1.4. *Independence models* generalize Example 2.1.3 in a natural way. Let Δ_1 be the 0-dimensional simplicial complex consisting only of the vertices:

$$\Delta_1 := \{i : i \in N\} \tag{2.13}$$

The exponential family $\mathcal{E}_{\Delta_1}$ is called the *independence model on n random variables*. A Markov basis of the independence model is easy to derive. In Section 2.2 we will see that it consists entirely of elementary circuits of degree 2. The marginal polytope of the independence model is a product $\prod_{i\in N} \Delta_{(d_i-1)}$ of simplices of dimensions $d_i - 1$.

EXAMPLE 2.1.5 (Simplicial complexes from graphs). There are natural ways to construct a simplicial complex from a graph G. A *graph model* of G is the exponential family of the 1-dimensional simplicial complex consisting of the vertices and edges of G. In this slightly unfortunate nomenclature, a *graphical model* of a graph G is the exponential family for the clique complex of G. Markov bases and marginal polytopes of these models have received widespread attention in algebraic statistics and topological combinatorics [HS02; DS03; WJ03; GMS06; Eng08].

EXAMPLE 2.1.6 (k-interaction models). The independence model allows no interaction between the units. A generalization of this idea is to allow only interaction between k units. In our framework this can be simply achieved by defining

$$\Delta_k := \{B \subseteq N : |B| \leq k\}\,, \tag{2.14}$$

the *k-uniform simplicial complex* . The hierarchical model $\mathcal{E}_{\Delta_k}$ is called the *k-interaction model*. Obviously this definition includes the independence model via Δ_1. These models are central in information geometry [AKN92; Ama01]. Already Δ_2 is interesting as its closure contains all maximizers of the multi-information function [AK06; MA03]. Algebraically these models are also interesting, Δ_2 gives the graph model of the complete graph, a model whose Markov basis remains unknown as soon as $n > 5$. Also their marginal polytopes, at least in the binary case, can be recognized to be known complicated objects. We discuss them shortly in Section 2.4.1. Exploiting symmetry to understand k-interaction models is surely an interesting endeavor.

A considerable part of the structure theory of hierarchical models was developed in the works of Sullivant et al. [HS02; DS03]. A natural idea is to decompose simplicial complexes into their elementary parts, and study those separately. A simplicial complex Δ is called *reducible* if there exist two complexes Δ_1, Δ_2 and a set $S \in \Delta$ such that

(1) $\Delta = \Delta_1 \cup \Delta_2$.
(2) $\Delta_1 \cap \Delta_2 = 2^S$.
(3) $2^S \neq \Delta_i$ for $i = 1, 2$.

The complex Δ is called *decomposable* if it is reducible with Δ_i being either decomposable themselves, or simplices. Thus among the graph models exactly the trees are reducible, while a triangle, the simplicial complex [12][23][31], is not reducible. Decomposable models are the "easy" ones among all hierarchical models. Their Markov bases can be explicitly described and consist of simple quadratic moves.

2.1.1. Parametrization of Hierarchical Models. In Section 2.1 we have defined hierarchical models parametrically. Each point in an open hierarchical model is given by a vector of parameters $c \in \mathbb{R}^d$. A characteristic feature of a parametrization like (2.9) is that each $p \in \mathcal{E}_A$ is defined by the "energy" $H(x) = \langle c, A_x\rangle$ of x. Because of the normalization Z_c, which cancels any constant term (as a function of x) that one might add to H, we are lead to the consideration of the rowspan of A_Δ, modulo the constant functions.

DEFINITION 2.1.7. Let $\mathbb{1}$ be the constant function $x \mapsto 1$ on $\mathcal{X}$. The linear space

$$\mathcal{T}_\Delta := \operatorname{rowspan} A_\Delta / (\mathbb{R}\mathbb{1}) \tag{2.15}$$

is called the *interaction space* or *tangent space* of the hierarchical model associated with Δ.

The elements of the interaction space $\mathcal{T}_\Delta$ are in bijection with points in the open hierarchical model $\mathcal{E}_\Delta$. We define the exponential and logarithm:

$$\begin{aligned} \exp : \mathcal{T}_\Delta &\to \mathcal{E}_\Delta \\ (H(x))_{x\in\mathcal{X}} + \mathbb{R}\mathbb{1} &\mapsto \left(\frac{\exp H(x)}{\sum_{y\in\mathcal{X}} \exp(H(x))} \right)_{x\in\mathcal{X}}, \end{aligned} \tag{2.16}$$

$$\begin{aligned} \log : \mathcal{E}_\Delta &\to \mathcal{T}_\Delta \\ (p(x))_{x\in\mathcal{X}} &\mapsto (\log p(x))_{x\in\mathcal{X}} + \mathbb{R}\mathbb{1}. \end{aligned} \tag{2.17}$$

It is clear that these maps are well-defined, surjective, and inverse to each other. We can change the parametrization by choosing different generating sets of the rowspan of A: To this end, let $\mathfrak{B} := \{b_i : i = 1, \ldots, d\}$ be any finite generating system of $\mathcal{T}_\Delta$. Each choice gives a different parameterization of the hierarchical model. The parametrization is called *identifiable* if $\mathfrak{B}$ is a basis. In this case the parameters can be uniquely reconstructed from H. Naturally each choice of a basis gives an (affinely) equivalent description of the marginal polytope; the convex hull of the columns of the matrix that has $\mathfrak{B}$ as its rows. Some choices will be discussed now.

Marginals: We have used this representation in the definition of a hierarchical model. It is a natural choice, coming from the analysis of contingency tables. The model is not identifiable and the number of parameters is typically too large. The representation of A as a 0/1-matrix however has numerous advantages, the most prominent being the applicability of toric algebra.

Statistical Physics - Potentials: In statistical mechanics one considers *potentials* [Win03; Geo88]. A potential is a collection of functions $U_B, B \subseteq N$, where U_B depends on its argument x only through $X_B(x)$, and $U_\emptyset = 0$, such that the energy can be written as a linear combination hereof. Often one has a distinguished state o called the vacuum. A potential is called *normalized* if $U_B(x) = 0$ as soon as $x_i = o_i$ for some $i \in B$. Given a strictly positive distribution, a corresponding normalized potential exists and is unique. For example in the binary setting, choosing $(0, 0, \ldots, 0)$ as the vacuum state, the normalized potential is given by the functions $U_B = c_B \prod_{i\in B} x_i$, where $c_B \in \mathbb{R}$. Then $\mathfrak{B} = \left\{\prod_{i\in B} x_i : B \in \Delta\right\}$ is a basis of the interaction space. Note that $B = \emptyset$ gives the constant function $x \mapsto 1$. Expanding a function $H \in \mathbb{R}^{\mathcal{X}}$ in terms of this basis was called the χ-expansion in the works of Caianiello [Cai75; Cai86].

An orthogonal basis of characters: In the binary case, where $\boldsymbol{d} = (2, 2, \ldots, 2)$, and thus $\mathcal{X} \cong \{0, 1\}^N$, a natural basis for $\mathbb{R}^{\mathcal{X}}$ is given by the characters of $\mathcal{X}$. Here, we assume coordinate-wise addition modulo 2 as the group operation. For every set $B \in \Delta$ define the function $e_B : \mathcal{X} \to \{-1, 1\}$ by

$$e_B(x) := (-1)^{E(B,x)} \tag{2.18}$$

where $E(B, x) := |\{i \in B : x_i = 1\}|$. It can be seen that, if Δ is a simplicial complex, $\{e_B : B \in \Delta\}$ is an orthogonal basis of the rowspan of A_Δ. If we treat $\mathcal{X}$ as the additive group $(\mathbb{Z}/2\mathbb{Z})^n$ then the characters of this group form an orthonormal basis (with respect to the product induced by the Haar measure of $\mathbb{C}^{\mathcal{X}}$. In our case the Haar measure is proportional to the standard product of $\mathbb{C}^{\mathcal{X}}$. The characters are exactly given by the vectors $e_B, B \subseteq N$. They are real valued functions, and thus a basis of $\mathbb{R}^{\mathcal{X}}$ [Pon66; KA06].

Various people, starting with Caianiello [Cai75] have called this the η-expansion. Note that if one considers random variables taking values in $\{\pm 1\}$, this basis equals the monomial basis $\left\{\prod_{i\in B} x_i : B \subseteq [N]\right\}$ considered above.

A basis of parity functions Finally, we introduce yet another basis of the rowspace of A, which is derived from the basis of characters. To each $\emptyset \neq B \subseteq [N]$, we define a vector in $\mathbb{R}^{\mathcal{X}}$

$$f_B(x) := \begin{cases} 1 & \text{if } |\operatorname{supp}(x) \cap B| \text{ is odd} \\ 0 & \text{otherwise.} \end{cases} \tag{2.19}$$

Denote again $\mathbb{1} : \mathcal{X} \to \mathbb{R}$, the constant function $x \mapsto 1$. Since $e_B(x) = 1 - 2f_B(x)$ for any $B \neq \emptyset$ we have that $\{f_B : \emptyset \neq B \in \Delta\} \cup \{\mathbb{1}\}$ is a basis of the rowspan of A_Δ. One interesting fact about this representation is that it gives full-dimensional 0/1-polytopes, the vertices of which form an additive group and thereby a linear code (see Proposition 2.4.7). For all other choices of $\mathfrak{B}$ discussed in this section, this is not the case. In Section 2.4.2 we will study the consequences of this observation.

The monomial parameterization The aforementioned parameterizations reach only points interior to the probability simplex, giving distributions with full support. Limit distributions have to be described by appropriate compactification of $\mathbb{R}^d$, the space of parameters. This introduces various technical difficulties which seem avoidable by the following formal trick. We first rewrite Definition 2.9 in a multiplicative way:

$$\mathcal{E}_\Delta = \left\{ p \in \mathcal{P}(\mathcal{X}) : p(x) = \theta^{a_x}, \theta \in \mathbb{R}^d_{>0} \right\}. \tag{2.20}$$

where we used monomial notation $\theta^{a_x} := \prod_{i=1}^{k} \theta_i^{a_x(i)}$. The advantage of this parameterization is, that we can simply replace the space of parameters by $\mathbb{R}^d_{\geq 0}$, allowing zero, to include parts of the boundary of the model. However, it can be seen that the extended model is not necessarily the closure of the hierarchical model, and even worse, depends on the explicit choice of the matrix A. Nevertheless, a monomial parameterization of a geometric object is a very useful tool. It enables to compute implicit representations as discussed in Section 1.4.

2.2. Elementary Circuits of Hierarchical Models

In this section we study the kernels of matrices $A_{\Delta,\boldsymbol{d}}$ more closely. We give a class of inclusion minimal integer kernel elements, i.e. circuits, and study how these relate to Markov bases of hierarchical models. We have seen in Example 1.4.2 that the set of circuits need not include a Markov basis. However, it turns out that it is not so easy to find an example of a *hierarchical* model whose Markov basis does not consist entirely of circuits. In [AT03], S. Aoki and A. Takemura give a model and a Markov basis element which is not a circuit. Interestingly, the full Markov basis of this model is not known.

The elementary circuits generalize vectors that have been discussed in [Kah10b] and the adjacent minors of [HS02]. For the following let again $\Delta \subseteq 2^N$ be a simplicial complex. The state space is the product $\mathcal{X} = \prod_{i\in N} \mathcal{X}_i$. We frequently need to distinguish the binary case $\mathcal{X} = \{0,1\}^N$. Many arguments can be reduced to this case and it is also our main example.

Definition 2.2.1. Let $(G, y_{N\setminus G}, y_G^+, y_G^-)$ be a 4-tuple consisting of a set $G \subseteq N$, a configuration on the complement $N\setminus G$: $y_{N\setminus G} \in \mathcal{X}_{N\setminus G}$, and two configurations $y_G^+, y_G^- \in$

$\mathcal{X}_G$ on G, such that $y_i^+ \neq y_i^-$ for all $i \in G$. The *elementary circuit* corresponding to this data is the vector with components

$$c_G^{y_{N\setminus G},y_G^+,y_G^-}(x) := \begin{cases} (-1)^{\left|\left\{i\in G: y_i^- = x_i\right\}\right|} & \text{if } x_{N\setminus G} = y_{N\setminus G} \text{ and } x_i \in \left\{y_i^+, y_i^-\right\}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.21}$$

REMARK 2.2.2. The set of elementary circuits contains many proportional vectors as stated in the following proposition. Whenever possible we strive to simplify the notation and use c_G, meaning an arbitrary choice of $y_{N\setminus G}$ and y_G^+, y_G^-, i.e. an arbitrary elementary circuit for the nonface G.

PROPOSITION 2.2.3. *(i) Flipping a position of y^+ and y^- changes the sign. Formally, if for some $i \in G$, $z_i^+ = y_i^-$ while $z_j^+ = y_j^+$ for $j \neq i$, and vice versa $z_i^- = y_i^+$ while $z_j^- = y_j^-$ for $j \neq i$, then we have $c_G^{y_{N\setminus G},z^+,z^-} = -c_G^{y_{N\setminus G},y^+,y^-}$.*
(ii) For each nonface $G \notin \Delta$, the elementary circuits are kernel elements $c_G \in \ker A_\Delta$.
(iii) The degree of c_G equals $2^{|G|-1}$.

PROOF. *(i)* This property follows directly from the definition since the flip does not change the support of the corresponding circuits.
(ii) To show that $A_\Delta c_G = 0$ we check the product on an arbitrary row of A_Δ, given by a set $B \in \Delta$ and a configuration y_B. The scalar product of this row and c_G is $\sum_{x\in\{X_B=y_B\}} c_G(x)$. Now select $j \in G \setminus B$ and group terms according to the value of x_j:

$$\sum_{x\in\{X_B=y_B\}} c_G(x) = \sum_{\substack{x\in\{X_B=y_B\}\\ \cap\{X_j=y_j^+\}}} (-1)^{\left|\left\{i\in G: y_i^- = x_i\right\}\right|} + \sum_{\substack{x\in\{X_B=y_B\}\\ \cap\{X_j=y_j^-\}}} (-1)^{\left|\left\{i\in G: y_i^- = x_i\right\}\right|} \tag{2.22}$$

$$= \sum_{\substack{x\in\{X_B=y_B\}\\ \cap\{X_j=y_j^+\}}} (-1)^{\left|\left\{i\in G\setminus\{j\}: y_i^- = x_i\right\}\right|} - \sum_{\substack{x\in\{X_B=y_B\}\\ \cap\{X_j=y_j^-\}}} (-1)^{\left|\left\{i\in G\setminus\{j\}: y_i^- = x_i\right\}\right|}. \tag{2.23}$$

This quantity equals zero since there is a natural bijection between the terms in the two sums and the signs disagree.
(iii) This follows directly from the definition. □

The binomials corresponding to elementary circuits of degree two look like independence statements. The higher ones are mimicking this with "checkerboard" patterns whose degrees are also powers of two. Below, in Section 2.2.4, we clarify the relation between elementary circuits and independence statements. Variants of such vectors are encountered occasionally in applications. In [O'S09] elementary circuits of degree 2^{n-1}, corresponding to $G = N$ are called *checkerboard vectors*. The "basis of vertex dependencies" in [HS02] is simply a basis of $\ker A_\Delta$, consisting entirely of elementary circuits. The authors define a set $\beta(S)$ of "adjacent minors" supported on sets $S \subseteq N$. Elements of $\beta(S)$ are elementary circuits supported on $G = S$ in our terminology. Their Theorem 2.6 extracts a basis of $\ker A$ from the set of elementary circuits. The same moves are also considered in Section 5 of [HS04]. In the following we aim at a theorem which justifies the name "elementary circuits".

THEOREM 2.2.4. *For fixed simplicial complex Δ and cardinality vector $\boldsymbol{d}$, let G be inclusion-minimal among the nonfaces of Δ. Then the elementary circuits $c_G^{y_{N\setminus G},y_G^+,y_G^-}$ are circuits of $A_{\Delta,\boldsymbol{d}}$ for any $y_{N\setminus G} \in \mathcal{X}_{N\setminus G}$, $y_G^\pm \in \mathcal{X}_G$.*

The proof of this theorem needs some additional machinery to enter the scene. It proceeds by reduction to the binary case $d = (2, 2, \ldots, 2)$, the idea being that a counterexample in the nonbinary case provides one in the binary case too. We first look at the binary case explicitly and then discuss a construction called *collapsing*.

2.2.1. Binary Elementary Circuits. In the following we denote $\mathcal{X} := \{0,1\}^N$. Put $\Delta^c := 2^N \setminus \Delta$ the set of nonfaces of Δ. For elements $G \in \Delta^c$ we define the upper intervals

$$[G, N] := \{B \subseteq N : B \supseteq G\} \subseteq \Delta^c \tag{2.24}$$

For each $B \subseteq N$ we have the vector $e_B \in \mathbb{R}^{\mathcal{X}}$, defined in Section 2.1.1. As discussed there, it is not difficult to see that $\{e_B : B \subseteq N\}$ is an orthogonal basis of $\mathbb{R}^{\mathcal{X}}$ such that $\{e_B : B \in \Delta^c\}$ is a basis of $\ker_{\mathbb{Z}} A_\Delta$. In the binary case, up to a sign, there is only one elementary circuit for each choice of $(G, y_{N\setminus G})$: $c_G^{y_{N\setminus G}}$. In the following denote $\boldsymbol{0} := (0, \ldots, 0)$.

LEMMA 2.2.5. *Let $G \in \Delta^c$, for $g := |G|$ it holds*

$$c_G^{\boldsymbol{0}}(x) = 2^{g-n} \sum_{B \in [G,N]} e_B(x) = \begin{cases} e_G(x_G) & \text{if } x_{N\setminus G} = \boldsymbol{0}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.25}$$

Furthermore, for any $C \subseteq N$, and $x_C \in \mathcal{X}_C$, we have the identity

$$2^{|C|-n} \sum_{x \in \{X_C = y_C\}} e_B(x) = \begin{cases} e_B(y_C) & \text{if } B \subseteq C, \\ 0 & \text{otherwise.} \end{cases} \tag{2.26}$$

PROOF. For the second case in (2.25) assume we have $i \in N \setminus G$ such that $x_i = 1$. Since half of the sets in $[G, N]$ contain i, while the other half does not contain i, it follows that the sum equals zero if such an i exists. The first case is now clear: all the summands are equal to e_G, and there are exactly 2^{n-g} terms. The identity (2.26) follows by the same argument. □

By choosing appropriate signs in the sum (2.25), one can achieve any of the binary elementary circuits, supported on cylinder sets $\{X_{N\setminus G} = x_{N\setminus G}\}$ instead of $\{X_{N\setminus G} = \boldsymbol{0}\}$. More concretely, we have

$$\begin{aligned} c_G^{y_{N\setminus G}}(x) &:= 2^{g-n} \sum_{B \in [B,N]} (-1)^{E(B, y_{N\setminus G})} e_B(x), \\ &= \begin{cases} e_G(x_G) & \text{if } x_{N\setminus G} = y_{N\setminus G} \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{2.27}$$

In Proposition 2.2.7, it follows that choosing G minimal in Δ^c, the value $2^n - 2^{|G|}$, as in Lemma 2.2.5, is the maximal number of zero components that can be achieved by nontrivial linear combinations of the vectors $e_B, B \in \Delta^c$, and thus by any vector in $\ker A$. Moving towards a proof of Theorem 2.2.4 in the binary case, we deduce a technical, but elementary statement about large subsets of $\mathcal{X}$.

LEMMA 2.2.6. *Let $g \in \{1, \ldots, n\}$ be fixed. For $\mathcal{Y} \subseteq \mathcal{X}$ with $|\mathcal{Y}| > 2^n - 2^g$ the following statement holds:*

- *For each $B \subseteq N$ with $|B| \geq g$, $\mathcal{Y}$ contains one of the cylinder sets $\{X_B = x_B\}$. More formally: $\exists x_B \in \mathcal{X}_B$ such that $\{X_B = x_B\} \subseteq \mathcal{Y}$.*

PROOF. The statement follows from a simple cardinality argument. Assume the contrary, let B be given, and $\forall x_B \in \mathcal{X}_B$, $\exists x \in \mathcal{X} \setminus \mathcal{Y}$ such that $x_B = X_B(x)$. These x are all distinct, since they differ on B. We find $|\mathcal{Y}| \leq 2^n - 2^g$. $\square$

PROPOSITION 2.2.7. *Let g denote the minimal cardinality among the sets in Δ^c. Then any nonzero linear combination of the vectors $e_B, B \in \Delta^c$ has at least 2^{g-1} positive and 2^{g-1} negative components.*

PROOF. Assume we have a linear combination

$$m = \sum_{B \in \Delta^c} z^B e_B \in \ker A_\Delta \tag{2.28}$$

which has less then 2^{g-1} positive components. It has at least $2^n - 2^{g-1} + 1$ nonpositive components. Let $\mathcal{Y}_\leq \subseteq \mathcal{X}$ denote the corresponding indices. Let $G \in \Delta^c$ have cardinality g and choose $i \in G$ arbitrary. By Lemma 2.2.6 we find a cylinder set $\left\{X_{G\setminus\{i\}} = y_{G\setminus\{i\}}\right\}$ that is contained in $\mathcal{Y}_\leq$. We have

$$m(x) = \sum_{B \in \Delta^c} z^B e_B(x) \leq 0 \qquad x \in \mathcal{Y}_\leq. \tag{2.29}$$

Summing up these equations over the cylinder set $\left\{X_{G\setminus\{i\}} = y_{G\setminus\{i\}}\right\}$ yields

$$\sum_{x \in \left\{X_{G\setminus\{i\}} = y_{G\setminus\{i\}}\right\}} \sum_{B \in \Delta^c} z^B e_B(x) \leq 0. \tag{2.30}$$

Note that this summation is in fact the computation of the marginal $m_{G\setminus\{i\}}$ evaluated at the value $y_{G\setminus\{i\}}$. Since $m \in \ker A_\Delta$, and $G \setminus \{i\} \in \Delta$, equality must hold in (2.30). We find that every term in the sum was already zero:

$$\sum_{B \in \Delta^c} z^B e_B(x) = 0 \qquad x \in \left\{X_{G\setminus\{i\}} = y_{G\setminus\{i\}}\right\}. \tag{2.31}$$

We inductively show that $m = 0$. Contained in $\left\{X_{G\setminus\{i\}} = y_{G\setminus\{i\}}\right\}$ we have a smaller set $\{X_G = y_G\}$. Summing up the respective components of m for this set, and using Lemma 2.2.5, we find

$$\begin{aligned} 0 &= \sum_{x \in \{X_G = y_G\}} \sum_{B \in \Delta^c} z^B e_B(x) \\ &= z^G 2^{n-g} e_G(x_G). \end{aligned} \tag{2.32}$$

It follows that $z^G = 0$. Applying the same argument, we can show that all coefficients z^H vanish for $|H| = g$. Inductively, we continue with sets of cardinality $g + 1$. Finally, this argument yields that all coefficients vanish and m is zero. The whole procedure applies, mutatis mutandis, for the negative components as well. $\square$

Now we prove that elementary circuits are circuits in the sense of matroid theory.

PROOF OF THEOREM 2.2.4 IN THE BINARY CASE. Let G be an inclusion minimal nonface. Fix a vector $c := c_G^{y_{N\setminus G}}$ and consider the submatrix $\tilde{A}$ whose columns are exactly those columns of A_Δ that correspond to the support of c. The statement is that this matrix has a single circuit consisting of all its columns. For each $i \in G$, by minimality of G, each $G \setminus \{i\}$ is contained in a facet of Δ. We can assume that $\tilde{A}$ has all rows computing marginals corresponding to $G \setminus \{i\}$ for $i \in G$, as these rows are linear combinations of the rows for the facet containing the respective set. It follows that $\tilde{A}$ has a submatrix which is the marginal matrix of the hierarchical model for the simplicial complex $2^G \setminus \{G\}$. By Proposition 2.2.7 this model's circuits have full support. $\square$

We next explain the collapsing construction, which is essential for reducing arguments to the binary case.

2.2.2. Collapsing. The idea of the following construction is to identify levels of each $\mathcal{X}_i$ until it is binary. Let $\mathcal{X} = \prod_{i=1}^n \mathcal{X}_i$ denote the nonbinary state space and additionally $\mathcal{Y} := \{0,1\}^n$.

DEFINITION 2.2.8. Let $\phi_i : \mathcal{X}_i \to \{0,1\}, i \in N$, be surjective maps. For each $B \subseteq N$, the composed maps

$$\begin{aligned} \phi_B : \mathcal{X}_B &\to \mathcal{Y}_B, \\ x_B &\mapsto (\phi_i(x_i))_{i\in B}. \end{aligned} \tag{2.33}$$

are called *collapsing maps*. We have an induced map on real valued functions:

$$\begin{aligned} \phi : \mathbb{R}^{\mathcal{X}} &\to \mathbb{R}^{\mathcal{Y}}, \\ (u(x))_{x\in\mathcal{X}} &\mapsto \left(\sum_{w\in\phi_N^{-1}(z)} u(w) \right)_{z\in\mathcal{Y}}. \end{aligned} \tag{2.34}$$

By considering these maps for each facet of Δ we have an induced map on margins too:

$$\begin{aligned} \Phi : \bigoplus_{F\in\Delta} \mathbb{R}^{\mathcal{X}_F} &\to \bigoplus_{F\in\Delta} \mathbb{R}^{\mathcal{Y}_F}, \\ (u_F)_{F\in\Delta} &\mapsto (\phi(u_F))_{F\in\Delta}. \end{aligned} \tag{2.35}$$

The key property of a collapsing is that it commutes with marginalization:

LEMMA 2.2.9. *For fixed simplicial complex Δ denote $A_{\boldsymbol{d}}$ the marginal matrix for levels $\boldsymbol{d}$ and A_2 the binary one. The following diagram commutes:*

$$\begin{tikzcd} \mathbb{R}^{\mathcal{X}} \arrow[r, "A_d"] \arrow[d, "\phi"] & \bigoplus_{F\in\mathcal{F}} \mathbb{R}^{\mathcal{X}_F} \arrow[d, "\Phi"'] \\ \mathbb{R}^{\mathcal{Y}} \arrow[r, "A_2"'] & \bigoplus_{F\in\mathcal{F}} \mathbb{R}^{\mathcal{Y}_F} \end{tikzcd} \tag{2.36}$$

PROOF. Let $u \in \mathbb{R}^{\mathcal{X}}$. We claim that for an arbitrary component of a vector in $\bigoplus_{F\in\mathcal{F}} \mathbb{R}^{\mathcal{Y}_F}$, defined by $B \subseteq N, z_B \in \mathcal{Y}_B$, it holds:

$$\begin{aligned} \Phi(A_d(u))(B, z_B) &= \sum_{x_B\in\phi_B^{-1}(z_B)} \sum_{w\in\{X_B=x_B\}} u(w), \\ &= \sum_{y\in\{X_B=z_B\}} \sum_{w\in\phi^{-1}(y)} u(w) \\ &= A_2(\phi(u))(B, z_B). \end{aligned} \tag{2.37}$$

Note that for the cylinder set on the left hand side, $\{X_B = x_B\} \subseteq \mathcal{X}$, while on the right hand side $\{X_B = z_B\} \subseteq \mathcal{Y}$. Since on each side every w appears at most once, it suffices to show the equality of sets

$$\bigcup_{x_B\in\phi_B^{-1}(z_B)} \{X_B = x_B\} = \bigcup_{y\in\{X_B=z_B\}} \left\{\phi^{-1}(y)\right\}. \tag{2.38}$$

"$\subseteq$": Let w from the left hand side be given. One has $X_B(w) = x_B$ for some x_B with $\phi_B(x_B) = z_B$. Therefore $\phi(w) = y$ with $X_B(y) = z_B$ and w is contained in the right hand side.

"$\supseteq$": Let $w \in \phi^{-1}(y), y \in \{X_B = z_B\}$ from the right hand side be given. We have $X_B(w) \in \phi_B^{-1}(z_B)$, so w is contained in the left hand side.

□

Next, for completeness, we have a quick look at the inverse operation of embedding binary tables into higher dimensional tables. Naturally one can choose "simple" inverses of collapsings. For some $B \subseteq N$, let ϕ_B be a collapsing. We can choose an injection $\lambda_B : \{0,1\} \to \mathcal{X}_B$ such that $\phi_B \circ \lambda_B = \mathrm{id}_{\{0,1\}^B}$. We get a map $\lambda : \mathbb{R}^{\mathcal{Y}} \to \mathbb{R}^{\mathcal{X}}$ by setting

$$\lambda(u)(x) = \begin{cases} u(y) & \text{if } y \in \operatorname{im} \lambda_N, \\ 0 & \text{otherwise,} \end{cases} \tag{2.39}$$

which again induces a map Λ on marginals. Consider for an example the embedding of a 2×2 table into a 3×3 table.

$$\lambda \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 0 \\ 3 & 4 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{2.40}$$

It is clear that one can compute the margins on the 2×2 table and fill them with zero's or compute the margins of the 3×3 table. This is the content of the following

LEMMA 2.2.10. *We have $\Lambda A_2 = A_d \lambda$, and the following diagram commutes:*

$$\begin{tikzcd} \mathbb{R}^{\mathcal{X}} \arrow[r, "A_d"] & \bigoplus_{F \in \mathcal{F}} \mathbb{R}^{\mathcal{X}_F} \\ \mathbb{R}^{\mathcal{Y}} \arrow[u, "\lambda"] \arrow[r, "A_2"'] & \bigoplus_{F \in \mathcal{F}} \mathbb{R}^{\mathcal{Y}_F} \arrow[u, "\Lambda"'] \end{tikzcd} \tag{2.41}$$

PROOF. By permutation of labels in the $\mathcal{X}_i$, we can assume that λ satisfies $\lambda(0) = 0$ and $\lambda(1) = 1$. Again, let $B \in \Delta$ and $x_B \in \mathcal{X}_B$ be fixed.

$$\begin{aligned} A_{\boldsymbol{d}}(\lambda(u))(x_B) &= \sum_{x \in \{X_B = x_B\}} \lambda(u)(x_B) \\ &= \begin{cases} \sum_{x \in \{X_B = x_B\}} u(x_B), & \text{if } x_B \in \{0,1\}, \\ 0, & \text{otherwise} \end{cases} \\ &= \Lambda(A_2(u))(x_B). \end{aligned} \tag{2.42}$$

□

2.2.3. The Nonbinary Case.

PROPOSITION 2.2.11. *Given a nonzero $c \in \ker A_{\Delta,\boldsymbol{d}}$ with $|supp(c)| = s$, there exists a nonzero $c' \in \ker A_{\Delta,\boldsymbol{2}}$ with $\mathrm{supp}(c') \leq s$. In particular Proposition 2.2.7 holds also in the nonbinary case.*

PROOF. The statement follows by application of a proper collapsing. There exists a collapsing ϕ such that $c' := \phi(c)$ is not zero. The support of c' has smaller cardinality. Now, if there was a violation of Proposition 2.2.7, this construction would give a counterexample in the binary case. □

Proof of Theorem 2.2.4. We reduce the statement to the binary case. By definition, for each elementary circuit c of the model associated to $(\Delta, \boldsymbol{d})$ there exists a collapsing ϕ, such that $\phi(c)$ is an elementary circuit of the binary model associated to Δ. If $\tilde{c} \in \ker A_{\Delta,\boldsymbol{d}}$ had $\operatorname{supp}(\tilde{c}) \subsetneq \operatorname{supp}(c)$, then $\phi(\tilde{c}) \in \ker A_{\Delta,\boldsymbol{2}}$ would satisfy $\operatorname{supp}(\phi(\tilde{c})) \subsetneq \operatorname{supp}(\phi(c))$, which is impossible by the binary version of the theorem. □

Elementary circuits corresponding to sets of cardinality two correspond to independence statements. In the next section we will further look into this correspondence. It is shown that the ideal of the pairwise Markov condition is generated by elementary circuits of degree two. Then we will see that a quadratic generator in a hierarchical model always originates from an independence statement, but not necessarily a pairwise one. In a certain sense, the higher elementary circuits are generalized independence statements.

2.2.4. Independence Statements and Quadratic Generators. We begin by recalling graphical models and their relation to hierarchical models. The standard reference on the subject is [Lau96]. The new book [DSS09] exposes the algebraic viewpoint.

Definition 2.2.12. Let G be an undirected graph. The *graphical model* of G is the hierarchical model of the clique complex of G, that is the simplicial complex whose faces are the complete subgraphs of G.

Apart from the graphical model, a graph also induces conditional independence models. We need some further definitions for this.

Definition 2.2.13. A *conditional independence* $X_A \perp\!\!\!\perp X_B \,|X_C$ is said to hold for a distribution p if its $A \cup B \cup C$ marginal p^{ABC} satisfies

$$p^{ABC}_{x_A x_B x_C} p^{ABC}_{y_A y_B x_C} = p^{ABC}_{x_A y_B x_C} p^{ABC}_{x_A y_B x_C}, \tag{2.43}$$

for all $x_A, y_A \in \mathcal{X}_A, x_B, y_B \in \mathcal{X}_B$ and $x_C \in \mathcal{X}_C$. If $A \cup B \cup C = N$ we say that the statement is *saturated.* In this case the defining equations are binomial.

For a fixed statement defined by A, B, C it can be seen that the ideal defined through the equations (2.43) is a prime ideal. The proof follows directly from general facts on determinantal ideals [BV88]. In general, questions about saturated conditional independence statements are leading to interesting topics in determinantal rings, the reason being that (2.43) describes rank conditions on submatrices of the probability tensor $(p_x)_{x\in\mathcal{X}}$.

Now, by *imposing* conditional independence statements we can also specify a statistical model called a *conditional independence model.* Given a graph G, the *pairwise Markov condition* states that $X_i \perp\!\!\!\perp X_j \,\big|X_{[n]\setminus\{i,j\}}$ for each nonedge $\{i,j\} \notin E(G)$, while the *global Markov condition* states that $X_A \perp\!\!\!\perp X_B \,|X_C$ whenever C separates A and B in G. The pairwise (global) Markov ideal I_{pair} (I_{global}) is the ideal generated by the binomial equations of all statements defined by the graph. It is a sum of prime ideals, its variety, called the pairwise (global) conditional independence model, is typically not irreducible and thus primary decompositions of these ideals are very interesting for understanding conditional independence in general. The algebraic interpretation of the Hammerslay-Clifford-Theorem shows that the graphical hierarchical model defined above is the variety of the saturated ideal $\big(I_{\text{global}} : \big(\prod_{x\in\mathcal{X}} p_x\big)^\infty\big)$, which equals $\big(I_{\text{pair}} : \big(\prod_{x\in\mathcal{X}} p_x\big)^\infty\big)$. The remaining primary components of I_{pair} and I_{global} all contain monomials and their varieties consist of distributions with limited support. When we restrict ourselves to distributions with full support, the Hammersley-Clifford-Theorem relates the different Markov conditions and the property of lying in the image of the parameterizations of the graphical model that we discussed in Section 2.1.1. It states that a strictly positive distribution lies in the image of the parameterization of the hierarchical (graphical) model of G, if and only if it satisfies the pairwise Markov condition.

Coming back to the elementary circuits we can now observe that each elementary circuit belongs to a pairwise independence statement:

PROPOSITION 2.2.14. *For configurations $x_1, \dots, x_4$, a quadratic equation*

$$p_{x_1}p_{x_2} - p_{x_3}p_{x_4}, \tag{2.44}$$

corresponds to a pairwise independence statement $X_i \perp\!\!\!\perp X_j \,\big|X_{[n]\setminus\{i,j\}}$ iff it is one of the elementary circuits $c_{\{i,j\}}$. The pairwise Markov ideal is generated by elementary circuits.

PROOF. If i and j are singletons, the statement $X_i \perp\!\!\!\perp X_j \,\big|X_{N\setminus\{i,j\}}$ exactly gives the equations that have all the elementary circuits $c_{\{i,j\}}$ as their exponent vectors. $\square$

EXAMPLE 2.2.15 (Example 2.1.3 continued). The independence model $\mathcal{E}_1$, the graphical model of the empty graph on n vertices, is generated by elementary circuits. These however are not algebraically independent. Let $n = 3$, a quick computation with 4ti2 reveals a Markov basis, consisting of the 9 vectors

$$\begin{aligned} m_1 &= (0,0,0,0,1,-1,-1,1), & m_2 &= (0,0,1,-1,0,0,-1,1),\\ m_3 &= (0,1,-1,0,0,-1,1,0), & m_4 &= (0,1,0,-1,-1,0,1,0),\\ m_5 &= (0,1,0,-1,0,-1,0,1), & m_6 &= (1,-1,-1,1,0,0,0,0),\\ m_7 &= (1,-1,0,0,-1,1,0,0), & m_8 &= (1,-1,0,0,0,0,-1,1),\\ m_9 &= (1,0,-1,0,-1,0,1,0). \end{aligned} \tag{2.45}$$

This Markov basis is not unique. The number of Markov moves is typically far smaller than the number of circuits. For the model at hand there are 20 circuits, 12 of them are elementary. The move

$$c = (0,0,1,-1,-1,1,0,0), \tag{2.46}$$

is a circuit that does not appear in the Markov basis. Its binomial is algebraically dependent as

$$p^{c^+} - p^{c^-} = (p^{m_4^+} - p^{m_4^-}) - (p^{m_3^+} - p^{m_3^-}). \tag{2.47}$$

We now have a look at quadratic generators in hierarchical models in general. The converse of Proposition 2.2.14 is that every quadratic generator is coming from a global independence statement. The idea for this proof is essentially contained in [DS03] and the author was pointed at this fact by Seth Sullivant. The following statement was conjectured by Ignacio Ojeda in a talk he gave in Torino in May 2009.

THEOREM 2.2.16. *Let G be a graph and I_{global} the ideal generated by the global Markov property. Let $I_{\text{toric}} = \left(I_{\text{global}} : (\prod_{x\in\mathcal{X}} p_x)^\infty\right)$ be the corresponding toric ideal. Then I_{global} is generated by the quadrics in the toric ideal.*

PROOF. Because of the containment $I_{\text{global}} \subseteq I_{\text{toric}}$ it suffices to show that each quadratic generator of I_{toric} is contained in I_{global}. This can be done by showing that each quadratic generator corresponds to an actual global independence statement. For convenience we introduce tableau notation [HS02] for monomials. In this notation, the monomial p^u is represented by listing each $x \in \mathcal{X}$, $u(x)$ times. For example $p_{000}p_{110}p_{111}^2$ is written as the tableau

$$\begin{bmatrix} 000 \\ 110 \\ 111 \\ 111 \end{bmatrix}. \tag{2.48}$$

In this notation a quadratic generator is written as

$$\begin{bmatrix} i_1, \dots, i_n \\ i'_1, \dots, i'_n \end{bmatrix} - \begin{bmatrix} j_1, \dots, j_n \\ j'_1, \dots, j'_n \end{bmatrix}. \tag{2.49}$$

Since any graphical model preserves the column marginals in these moves we have $\{i_k, i'_k\} = \{j_k, j'_k\}$ for any index $k = 1, \dots, n$. A homogeneous quadratic binomial has at most 2 variables in its positive and negative support, respectively, and we can assume that the configurations $x \in \mathcal{X}$ supporting this binomial are binary. If not, a relabeling inside the $\mathcal{X}_i$ provides such a move. Among the occurring indices we potentially find tuples $(i, i') = (j, j') = (0, 0)$ or $(i, i') = (j, j') = (1, 1)$. Without loss of generality we can assume to find these in the beginning and the end, and write

$$\begin{bmatrix} 0, \dots, 0, i_l, \dots, i_m, 1, \dots, 1 \\ 0, \dots, 0, i'_l, \dots, i'_m, 1, \dots, 1 \end{bmatrix} - \begin{bmatrix} 0, \dots, 0, j_l, \dots, j_m, 1, \dots, 1 \\ 0, \dots, 0, j'_l, \dots, j'_m, 1, \dots, 1 \end{bmatrix}. \tag{2.50}$$

In the remaining tuples $(i_k, i'_k), (j_k, j'_k)$, by permuting 0 and 1, we can achieve $(i_k, i'_k) = (0, 1)$ and thus assume that the move is

$$\begin{bmatrix} 0, \dots, 0, 0, \dots, 0, 0 \dots, 0, 1, \dots, 1 \\ 0, \dots, 0, 1, \dots, 1, 1 \dots, 1, 1, \dots, 1 \end{bmatrix} - \begin{bmatrix} 0, \dots, 0, 0, \dots, 0, 1, \dots, 1, 1, \dots, 1 \\ 0, \dots, 0, 1, \dots, 1, 0, \dots, 0, 1, \dots, 1 \end{bmatrix}. \tag{2.51}$$

Now, among the middle indices call the conserved pairs $A \subseteq N$, the flipped ones $B \subseteq N$, and the rest $C \subseteq N$, then we deduce that in this graphical model there is no edge between A and B since pair margins are not conserved. But in terms of conditional independence this just means that $A \perp\!\!\!\perp B \,|C$. The move is one of the generators coming from this statement and thus contained in I_{global}. ☐

Note that the proof works for any hierarchical model. It shows that a quadratic generator always corresponds to an independence statement holding in the model. This should be compared to the statement of Theorem 2.3.5: A hierarchical model for a simplicial complex containing all two element sets has no quadratic generators, as it has no valid independence statements. In the case of binary graph models, a generalization of this is the characterization of fixed-degree moves using elementary graphs in [DS03].

2.3. Markov Bases

As discussed in Chapter 1, if the kernel of A is generated by integer vectors, we are in the toric case and can apply commutative algebra. Hierarchical models fall into this class; A_Δ is a 0/1-matrix. The toric ideal is

$$I_\Delta := I_{A_\Delta} = \langle p^u - p^v : A_\Delta u = A_\Delta v,\ u, v \in \mathbb{N}^d \rangle. \tag{2.52}$$

The nonnegative real part of the variety of I_Δ is the closure of the hierarchical model: $V_{\geq 0}(I_\Delta) = \overline{\mathcal{E}_\Delta}$. In this language, a *Markov basis* is a finite set of vectors $M \subseteq \mathbb{Z}^d$, such that

$$I_\Delta = \langle p^{m^+} - p^{m^-} : m \in M \rangle. \tag{2.53}$$

Markov bases of hierarchical models are rich in structure. The new text book [DSS09] discusses them along with other "bases" of integer lattices. We try to avoid repeating too much of the theory here, but restrict ourselves to new results related to elementary circuits. To this end, in Section 2.3.1, we discuss the class of models whose Markov bases consist precisely of elementary circuits. We show that in this case the matrix A is already totally unimodular. This result generalizes a theorem in [Kah10b] and appears here for the first time. In Section 2.3.2 we show corollaries of our discussion on the boundedness of support size.

During the work on this thesis the author had to juggle a lot of computational results. To facilitate this experimental research and make it available for the public, an Internet database was set up by Johannes Rauh and the author. In Section 2.3.3 we describe MBDB, the Markov Bases Database.

2.3.1. Total Unimodularity of Complement Complexes. Elementary circuits generalize independence statements. Parallel to the independence model, it is natural to ask for a class of models, characterized by these moves. We describe this class, and show that their marginal matrices are totally unimodular. This implies that the elementary circuits form a Markov basis, and even better, also coincide with each Gröbner and the so called Graver basis.

A Gröbner basis of a lattice can simply be defined as the set of exponent vectors of a Gröbner basis of the corresponding toric ideal. A *Graver basis* $\mathcal{G}$ of a lattice $\mathcal{L}$ is the unique minimal subset of $\mathcal{L}$ such that every $v \in \mathcal{L}$ has a *sign consistent* representation

$$v = \sum_{g \in \mathcal{G}} \lambda_g g, \quad \lambda_g \in \mathbb{N}, \qquad \text{where } |v_i| = \sum_{g \in \mathcal{G}} \lambda_g |g_i| . \tag{2.54}$$

The book [DSS09] contains many details and also the lattice interpretation of a Gröbner basis.

Definition 2.3.1. A matrix $A \in \mathbb{Z}^{m \times n}$ is called *totally unimodular* if each of its subdeterminants, and thus also each entry, is zero or ± 1.

It is often nontrivial to prove total unimodularity. Here we rely on the following lemma, variants of which are probably very well known in combinatorial optimization literature.

Lemma 2.3.2. *Let A be a 0/1-matrix such that:*

- *Each row of A has at most two nonzero entries.*
- *The set $\{a_1, \ldots, a_m\}$ of columns of A has a bipartition into sets M, N such that no two ones of a given row are contained in one of the sets M, N.*

In this case A is totally unimodular.

Proof. Without loss of generality we can assume that A is a square matrix. The proof is by induction on the dimension of the matrix and the 1×1 or 2×2 matrices are easily checked. Assume the statement holds for dimension $(n-1)$ and consider any $n \times n$ matrix A, satisfying the assumptions. If A has a zero row, the statement follows, and if A has a row with exactly one entry 1, we can expand on that column and use the induction hypothesis. Now, if every row of A has exactly two entries 1, then the columns a_i of A satisfy

$$\sum_{i \in M} a_i = \sum_{j \in N} a_j, \tag{2.55}$$

follows and thus $\det(A) = 0$. □

Let $G \subseteq N$. We denote

$$\Delta_{/G} := \{B \subseteq N : B \not\supseteq G\}, \tag{2.56}$$

the simplicial complex of all sets not containing G. We have seen that the toric ideal for this complex is generated in degree at least $2^{|G|-1}$. We now show that if Δ has the structure (2.56), and some restriction on $\boldsymbol{d}$ are fulfilled, the toric ideal is generated in degree exactly $2^{|G|-1}$, namely by the elementary circuits. The following result is a generalization of a theorem appearing in [Kah10b].

THEOREM 2.3.3. *Let $G \subseteq N$ and $d_i = 2$ for $i \in G$. The marginal matrix $A_{\Delta_{/G},\boldsymbol{d}}$ is totally unimodular. The set of elementary circuits*

$$M := \left\{ c_G^{y_{N\setminus G}} : y_{N\setminus G} \in \mathcal{X}_{N\setminus G} \right\}. \tag{2.57}$$

is the complete set of circuits and thus also forms a Markov, Gröbner, and Graver basis of $\ker A_{\Delta_{/G}}$.

REMARK 2.3.4. The investigation of the complexity of Markov bases of the no-three-way interaction model shows that varying the cardinalities $d_i, i \in G$ makes the statement fail. Markov bases of these models are "arbitrarily complicated" [LO06].

PROOF. We show that under the assumptions we have made, A is totally unimodular. The columns of A are indexed by $\mathcal{X}$ while each row is indexed by a facet $N \setminus \{i\}$, for some $i \in G$, and a configuration $x_{N\setminus\{i\}}$. Each row has entries 0 everywhere except in the columns $(x_{N\setminus\{i\}}, 0)$, $(x_{N\setminus\{i\}}, 1)$, which uses the hypothesis $d_i = 2$ for $i \in G$.

We bipartition the columns according to the parity (mod 2 sum of the entries) of x_G. According to Lemma 2.3.2 it remains to show that for each row the two nonvanishing entries lie in columns with different parity on G, but this is true since the two columns correspond to configurations that differ only in one entry x_i for $i \in G$, and thus their parity is different. This shows that A is totally unimodular and the set of circuits and the different bases coincide. We finish the proof by showing that the elementary circuits are actually all circuits in this case. We first compute the rank of A from the formula

$$\operatorname{rk} A = \sum_{F\in\Delta} \prod_{i\in F} (d_i - 1) \tag{2.58}$$

$$= |\mathcal{X}| - \sum_{F\notin\Delta} \prod_{i\in F} (d_i - 1), \tag{2.59}$$

which is a crucial ingredient for computing dimensions of discrete exponential families [Kah06; HS02]. In (2.59) we can change the summation because $d_i = 2$, for $i \in G$ and find

$$\sum_{F\supseteq G} \prod_{i\in F} (d_i - 1) = \sum_{F\subseteq N\setminus G} \prod_{i\in F} (d_i - 1) \tag{2.60}$$

$$= \begin{cases} 1, & \text{if } G = N, \\ \left|\mathcal{X}_{N\setminus G}\right|, & \text{otherwise,} \end{cases} \tag{2.61}$$

where the second equation follows from the rank formula again. Note that $\left|\mathcal{X}_{N\setminus G}\right|$ is the number of elementary circuits of A and their supports are disjoint. Any basis of the column matroid of A has to avoid $\left|\mathcal{X}_{N\setminus G}\right|$ columns, one from each circuit, as otherwise it would need to contain a circuit.

We claim that any maximal choice of columns not containing an elementary circuit gives a basis. Certainly one of them gives a basis $B = \{b_1, \ldots, b_n\}$. If b_i is an arbitrary element of this basis, it is contained in exactly one elementary circuit c as these are disjoint. On the other hand c contains exactly one element b', say, not contained in B. We claim that

$$B' = \{b_1, \dots, b_{i-1}, b', b_{i+1}, \dots, b_n\} \tag{2.62}$$

is again a basis. Indeed, if it was dependent, it would contain a circuit c', which must contain b', as otherwise B was not a basis. Now from the two circuits c and c' we can eliminate b' to find a third circuit consisting only of elements in B, a contradiction. This argument shows that any collection of $\operatorname{rk} A$ vectors avoiding all elementary circuits is independent and thus a basis. It also follows that the elementary circuits are the complete set of circuits of A, since if there was a bigger circuit not containing any of them, it would be contained in a basis. □

2.3.2. A Lower Degree Bound on Markov Moves. As a concluding observation on elementary circuits and lower degree bounds, we study the implications of Proposition 2.2.7 on Markov moves. Immediately we have the following

THEOREM 2.3.5. *Let Δ be a simplicial complex on N and I_Δ the corresponding toric ideal. Let g be the minimal cardinality of a nonface of Δ. Each generator of I_Δ has degree at least 2^{g-1}. Moreover, the positive and negative supports of each generator both have cardinality bigger or equal to 2^{g-1}. The degree bound is realized only by square-free binomials.*

This theorem gives a lower bound on the–smallest–degree among the generators. Lower bounds on the largest degree have been considered for a measure of complexity of the model for instance in [GMS06]. There it is shown that one finds a simplicial complex on $2n$ units, such that there exists a generator of degree 2^n. Furthermore, in [DS03] the authors study an algorithm that, for graph models, computes all generators of a given degree. Finally, in [LO06] the case of 2-margins of $(r, s, 3)$-tables is studied. It is shown that, as r and s grow, the supports and degrees of maximal generators cannot be bounded. This has interesting implications for data disclosure.

A typical example of hierarchical models are graph models where $\dim \Delta \leq 1$. If the graph Δ is not the complete graph, then the bound reduces to the trivial bound $\deg m \geq 2$. On the other hand, for the complete graph, there are no quadratic generators. This proves the interesting fact that on the hierarchical model $\mathcal{E}_2$ no global conditional independence statement is satisfied (see also Section 2.2.4).

2.3.3. The Markov Bases Database. In this final section on Markov bases we describe the Markov bases database (MBDB), available for public access on the web [RK]. It is a searchable collection of inputs and outputs of the program 4ti2. We explain the data structure and present examples of its usage. To begin with, each entry of the database lives inside its own directory, which has a semi-cryptic name such as "G114g_bin". These names are not very relevant and should just give distinguishable filenames. A verbal description of the model can be found inside the directory. The content of such a directory might look like the following list of files.

```
degreestats description G114g_bin.dot G114g_bin.mar
G114g_bin.mat G114g_bin.mod G114g_bin.png properties
```

They contain the following information

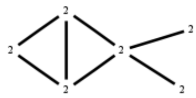
- **degreestats:** A list of degrees, and the number of generators in this degree, in a minimal Markov Basis.
- **description:** A verbal description of the model. This file reveals to us that here we find

```
The binary graphical model of G114.
```

G114g_bin.dot: A dot file describing the graph of this graphical model in the dot language [Tea].

G114g_bin.{mar,mat,mod}: The 4ti2 files containing the Markov basis, the marginal matrix and a 4ti2 readable description of the simplicial complex.

G114g_bin.png: An image of the model. G114 looks like this



The numbers on the vertices give the cardinalities of state space of the random variable at this site. As the description indicates, this model is binary.

properties: A list of properties that this model possesses or does not posses (with a leading "!")

Depending on the type of model and computations that were carried out, additional files, such as a .gro file containing a Gröbner basis, may be present.

To illustrate a use case of the database approach, we treat the following question using MBDB. It was posed as one of the exercises in Bernd Sturmfels' algebraic statistics course in Fall 2008 at Berkeley.

QUESTION 2.3.6. *Is there a hierarchical model that has a generator of odd degree? If yes, is there a binary model?*

The web interface offers some basic search possibilities, and the authors are always happy to hear about suggestions for new features. Other searches can be implemented after downloading a full dump of all data. Fortunately some statistics over degrees for each model is already available. A very simple call to the program `grep` solves our problem:

```
grep "^[3579] " */degreestats

BM5r3-5_bin/degreestats:9   32
G174_bin/degreestats:3   64
G174g_bin/degreestats:3   64
G174g_bin/degreestats:5   384
G189_bin/degreestats:5   768
G190_bin/degreestats:5   128
G196_bin/degreestats:5   512
G197_bin/degreestats:5   1920
G198_bin/degreestats:5   512
G199_bin/degreestats:5   1024
no3way-03-04-04/degreestats:9   576
no3way-03-04-05/degreestats:9   2880
no3way-04-04-04/degreestats:9   6912
```

The first line is the call. It asks `grep` to search for lines beginning with one of the numbers 3, 5, 7, or 9, followed by a space, in the files called `degreestats` in each subdirectory. Thus, this command has to be executed in the main data directory. Note that we only search for generators of the given degrees, but not of degree 11 and higher. The appropriate modification of the regular expression is not difficult. The output shows models that have generators of odd degree:

```
BM5r3-5_bin G174_bin G174g_bin G189_bin G190_bin
```

```
G196_bin G197_bin G198_bin G199_bin no3way-03-04-04
no3way-03-04-05 no3way-04-04-04
```

Some of them are binary. We have thus answered our question affirmatively.

2.4. Binary Marginal Polytopes

In the following we study special classes of marginal polytopes of hierarchical models. We have seen that knowledge about the face lattice of these polytopes is equivalent to knowledge on the support set problem. For instance Theorem 2.3.5 directly translates into a neighborliness property of marginal polytopes, given in Theorem 2.5.2. The starting point for observations in a different direction is the representation of binary marginal polytopes as 0/1-polytopes that was already outlined in Section 2.1.1. This allows to observe connections between marginal polytopes of graph models and CUT-polytopes. We establish the connection to polytopes that are convex hulls of linear codes. In Section 2.4.3 we classify all such polytopes and give an iterative algorithm which constructs them. This material was published in [KWA09].

2.4.1. Marginal, CUT, and correlation polytopes. Binary marginal polytopes are 0/1-polytopes, via the basis of parity function defined in Section 2.1.1. For any subset $B \subseteq N$, consider the vector $f_B \in \mathbb{R}^{\mathcal{X}}$ with components

$$f_B(x) := \begin{cases} 1 & \text{if } |\mathrm{supp}(x) \cap B| \text{ is odd,} \\ 0 & \text{otherwise.} \end{cases} \tag{2.63}$$

Consider the matrix whose rows are vectors $f_B, B \in \mathcal{D}$, for some collection $\mathcal{D} \subseteq 2^N$. We assign names to the columns of this matrix by defining

$$f_x^{\mathcal{D}}(B) := f_B(x). \tag{2.64}$$

Then f_x is a vector with $|\mathcal{D}|$ components. When the dependency on x is important we denote it as a function of x, writing $f^{\mathcal{D}}(x)$. If Δ is a simplicial complex, then the convex support of $\mathrm{cs}(\overline{\overline{\mathcal{E}_\Delta}})$ can be represented as $\mathrm{conv}\,\{f_B : B \in \Delta\}$. The definition of this polytope is interesting also in the general case of an arbitrary collection of vectors.

Definition 2.4.1. Let $\mathcal{D} \subseteq 2^N$ be a collection of sets. We define

$$\mathcal{F}_{\mathcal{D}} := \mathrm{conv}\left\{f^{\mathcal{D}}(x) : x \in \mathcal{X}\right\}. \tag{2.65}$$

These polytopes include binary marginal polytopes in the special case where $\mathcal{D}$ is a simplicial complex. We will now define the CUT-polytope of a graph, and show that if $\mathcal{D} = \{B \subseteq N, |B| \leq 2\}$ then $\mathcal{F}_{\mathcal{D}}$ is a CUT-polytope. Let $G = (V, E)$ be a graph. A *cut* $[A] = (G_1, G_2)$ is a bipartition of the vertex set of G, modulo the equivalence $(G_1, G_2) \sim (G_2, G_1)$, induced by taking complements. It follows that there are $2^{|V|-1}$ cuts. A *cut vector* for the cut $[A]$ is the 0/1-vector with components labeled by edges $e = (e_1, e_2) \in E(G)$:

$$\chi_{[A]}(e) := \begin{cases} 1 & \text{if } [A] \text{ cuts } e, \\ 0 & \text{otherwise.} \end{cases} \tag{2.66}$$

Definition 2.4.2. The CUT-polytope of a graph G is the convex hull of the cut vectors:

$$\mathrm{CUT}(G) := \mathrm{conv}\left\{\chi_{[A]} : [A] \text{ a cut in } G\right\}. \tag{2.67}$$

CUT-polytopes are central objects in combinatorial optimization, naturally arising in transportation problems. A book-length treatment of the subject is [DL97]. When the graph G is the complete graph K_n, the cut vectors have a very regular structure, resembling that of the vectors $f^{\mathcal{D}(x)}$. Namely, to each graph G we can associate the simplicial complex $\Delta_G = V(G) \cup E(G) \cup \{\emptyset\}$. We referred to this construction as *graph model* in Remark 2.1.5. From G construct the *coned graph* $\hat{G}$ with an additional vertex:

$$V(\hat{G}) := V(G) \cup \{*\}\,, \tag{2.68}$$

connected to all original vertices:

$$E(\hat{G}) := E(G) \cup \{(v,*) : v \in V(G)\}\,. \tag{2.69}$$

Then, denoting the CUT-polytope of $\hat{G}$ as $\mathrm{CUT}(\hat{G})$ one has

$$\mathcal{F}_{\mathcal{D}_G} = \mathrm{CUT}(\hat{G}). \tag{2.70}$$

Using the representation in terms of the vectors $f_{\mathcal{D}_G}(x)$, $x \in \mathcal{X}$, a proof of this equivalence becomes a renaming of coordinates. A cut in the coned graph is simply a subset of vertices of the original graph. In binary coding this translates to the binary string x. The coordinates of the ambient space are labeled by edges of the coned graph for $\mathrm{CUT}(\hat{G})$ and by subsets B with $|B| \leq 2$ for the marginal polytope. The bijection is the map that sends the edge to its set of vertices, intersected with $\{1, \ldots, n\}$. The following example should illustrate this.

Example 2.4.3. Consider the case $n = 2$ and let $\mathcal{D} = \{\{1\}, \{2\}, \{1,2\}\}$. The marginal polytope, represented with the vectors $f^{\mathcal{D}}$ of this model is the 3-simplex with coordinates

$$\begin{aligned} f(00) &= (0,0,0), \\ f(10) &= (1,0,1), \\ f(01) &= (0,1,1), \\ f(11) &= (1,1,0). \end{aligned} \tag{2.71}$$

The coned graph is a triangle, the full graph of 3 nodes. We order the coordinates as $\{1,2\}, \{1,3\}, \{2,3\}$. The distinct cuts can be labeled by $\emptyset, \{1\}, \{2\}, \{1,2\}$, resulting in cut vectors

$$\begin{aligned} \chi(\emptyset) &= (0,0,0), \\ \chi(\{1\}) &= (1,1,0), \\ \chi(\{2\}) &= (1,0,1), \\ \chi(\{1,2\}) &= (0,1,1). \end{aligned} \tag{2.72}$$

Remark 2.4.4 (Complexity). The realization of CUT-polytopes as marginal polytopes highlights the high complexity of this class of polytopes. In [DL97] it is shown that the problem of deciding whether a given point μ is an element of the CUT-polytope is NP-complete, as is enumerating the facets. Therefore one can not hope to find simple answers to decidability questions on the marginal polytope in general. If one restricts to subclasses, improvements are possible and the cited book (to a large extent) contains results in this direction.

Remark 2.4.5. The equivalence of marginal polytopes of graphical models, and CUT-polytopes of the corresponding coned graphs, is known to some extend in the literature on Markov random fields. For instance in [SJ08], constraints on the marginal polytope are found by iteratively projecting it to the CUT-polytope. This can be used to show upper bounds on the log-partition function. In [WJ03] the marginal polytope is discussed in connection to inference algorithms. Maximum-a-posteriori estimation can be seen as a linear program over the marginal polytope.

Concluding our section on different representations of marginal polytopes, based on the basis of products $\prod_{i\in B} x_i$ that was discussed in Section 2.1.1, we can find another class of known polytopes. Again, let $\mathcal{D}$ be any collection of sets. Let $B \subseteq N$ be arbitrary, we define the *inclusion vector* of B with respect to $\mathcal{D}$ as

$$c_{\mathcal{D}}(B) := (c_D(B))_{D\in\mathcal{D}} \quad \text{where} \quad c_D(B) := \begin{cases} 1 & \text{if } D \subseteq B, \\ 0 & \text{otherwise.} \end{cases} \tag{2.73}$$

The *inclusion polytope* is defined as the convex hull of all possible inclusion vectors.

$$\mathcal{C}_{\mathcal{D}} := \operatorname{conv}\{c_{\mathcal{D}}(B) : B \subseteq N\}. \tag{2.74}$$

The *correlation polytope* $\operatorname{COR}(N)$ of [DL97] is found included in this definition if we set $\mathcal{D} = \{B \subseteq N : |B| \le 2\}$. There exists an affine equivalence between $\operatorname{COR}(N)$ and $\operatorname{CUT}(K_{N+1})$ called the *covariance mapping*. It can be seen that this mapping generalizes to a mapping between binary marginal polytopes and the corresponding inclusion polytopes. In this sense, the parity representations $\mathcal{F}_{\mathcal{D}}$ of binary marginal polytopes form a generalization of CUT-polytopes to arbitrary simplicial complexes.

2.4.2. Binary Marginal Polytopes and Linear Codes. The starting point for this section is the simple observation that the vertices of binary marginal polytopes form a linear code when represented as in the previous section. To this end we make an additional assumption on the sets $\mathcal{D}$. From now on, $\mathcal{D}$ shall always contain all atoms $\{i\} \in \mathcal{D}$. Firstly we recall few definitions from coding theory. For a detailed introduction see for instance [van99]. The material presented here is joint work with Walter Wenzel and Nihat Ay and appeared in [KWA09].

Consider the finite field $\mathbb{F}_2 = (\{0,1\}, \oplus, \odot)$ with addition and multiplication mod 2. In coding theory, one studies particularly vector spaces over this field.

Definition 2.4.6. A binary $[r,s]$-*linear code* is a linear subspace L of $\mathbb{F}_2^r$ such that $\dim L = s$. Two codes are called *equivalent* if one can be transformed into the other by applying a permutation on the positions in the codewords, and for each position a permutation of the symbols. A *generator matrix* G for L is a $(s \times r)$-matrix that has as its rows a basis of L. Given L, one can find an equivalent code such that it has a generator matrix in *standard form*, i.e. $G = (E_s, H)$, where E_s is the $(s \times s)$-identity matrix.

The following proposition states that the vertices of $\mathcal{F}_{\mathcal{D}}$ form a linear code for any $\mathcal{D} \subseteq 2^N$. A special case of this connection has been mentioned in Example 2 of [WJ03]. In the following, whenever an ordering of $\mathcal{D}$ is required we silently assume that the atoms $\{i\} \in \mathcal{D}$ come first. This assumption gives generator matrices in standard form.

Proposition 2.4.7. *Let $\{0,1\}^{\mathcal{D}}$ be considered as a vector space over the finite field $\mathbb{F}_2$. Then the vertices $f^{\mathcal{D}}(x)$ form a linear subspace. If we consider $\mathcal{X} = \mathbb{F}_2^N$ as a vector space over $\mathbb{F}_2$, then the mapping $x \mapsto f^{\mathcal{D}}(x)$ is an injective homomorphism of vector spaces. Its image forms a $[|\mathcal{D}|, n]$-linear code, and a generator matrix in standard form has as its rows the vectors $f^{\mathcal{D}}_{e_i}$ for $i = 1, \ldots, N$, where e_i is the i-th unit vector in $\mathbb{F}_2^N$.*

Proof. The parity function is additive mod 2. Since scalar multiplication is trivial, we only need to show

$$f^{\mathcal{D}}(x \oplus y) = f^{\mathcal{D}}(x) \oplus f^{\mathcal{D}}(y) \qquad \text{for } x, y \in \mathcal{X}. \tag{2.75}$$

Let $B \in \mathcal{D}$, it suffices to show the identity for a single f_B. To do so, introduce

$$\begin{aligned} M &:= \{i \in B : (x_i = 1 \wedge y_i = 0) \vee (x_i = 0 \wedge y_i = 1)\}\,, \\ M_x &:= \{i \in B : x_i = 1\}\,, \\ M_y &:= \{i \in B : y_i = 1\}\,. \end{aligned} \tag{2.76}$$

Then $f_B(x \oplus y) = |M|$, $f_B(x) = |M_x|$, and $f_B(y) = |M_y|$. We find that M is the symmetric difference of M_x and M_y:

$$M = M_x \bigtriangleup M_y. \tag{2.77}$$

Since $|M_x \bigtriangleup M_y| = |M_x| + |M_y| - 2\,|M_x \cap M_y|$, we have that in $\mathbb{F}_2$

$$|M| = |M_x| \oplus |M_y|\,, \tag{2.78}$$

and therefore (2.75) holds. To see that $f^{\mathcal{D}}$ is injective, assume that $f_{\mathcal{D}}(x) = f_{\mathcal{D}}(y)$. Since $\mathcal{D}$ contains all atoms $\{i\} \subseteq N$, we get for every $i \in N$: $f_{\{i\}}(x) = f_{\{i\}}(y)$. This implies $x_i = y_i$, and hence, $x = y$. Since $\mathcal{X}$ considered as an $\mathbb{F}_2$ vector space has dimension n, also $f^{\mathcal{D}}(\mathcal{X})$ has dimension n and therefore forms an $[\mathcal{D}, n]$-linear code. $\square$

Every binary hierarchical model gives a linear code. In the following, we elaborate on a kind of reversal of Proposition 2.4.7. The points forming a linear code, when embedded into some Euclidean space, give an exponential family as in Chapter 1; they are a point configuration. When does there exist a hierarchical model defining the polytope at hand? This is the case "as soon as the number of parameters is right": Let $2^n \geq s \geq n$. Assume we are given an $[s, n]$-linear code. Without loss of generality, we assume that it has a generator matrix in standard form. We construct a collection $\mathcal{D}$ from the columns of the generator matrix. Since $\mathcal{D}$ is a set, while the columns are a list, repetitions of columns are lost. The codewords are given by the vertices of $\mathcal{F}_{\mathcal{D}}$. Let $E_n \in \mathbb{F}_2^{n\times n}$ denote the identity matrix in dimension n. Assume the generator matrix $G = (E_n, H) \in \mathbb{F}_2^{n\times s}$ has no two identical columns, implying $s \leq 2^n$. Denote by $\{e_i : i = 1, \ldots, n\}$ the canonical basis of $\mathbb{F}_2^n$. Using the columns of H, we define sets

$$B_j := \{i \in [n] : H_{ij} = 1\}\,, \qquad j = 1, \ldots, s - n \tag{2.79}$$

and then,

$$\mathcal{D} := \{\{1\}, \ldots \{n\}, B_1, \ldots, B_{s-n}\}\,. \tag{2.80}$$

Note that the elements of $\mathcal{D}$ are numbered in a natural way such that we can use $\mathcal{D}$ as an index set for the columns of $G = (G_{i,B})_{i=1,\ldots,n,\; B=\{1\},\ldots,\{n\},B_1,\ldots,B_{s-n}}$.

To see that $\left\{f^{\mathcal{D}}(e_i) : i = 1, \ldots, n\right\}$ is the set of rows of the generator matrix we evaluate

$$f_B(e_i) = \delta_{\{i \in B\}} = G_{i,B}, \tag{2.81}$$

which holds by definition of the B_j. Summarizing, every binary linear code (in standard form) corresponds to a collection $\mathcal{D} \subseteq 2^N$. However, two codes that differ only in repetitions of columns in the generator matrix are mapped to the same collection. Then, if $\mathcal{D}$ is a simplicial complex, the linear code is the linear code of a binary hierarchical model.

2.4.3. Classification of full-dimensional code polytopes. As we have seen, the polytopes $\mathcal{F}_{\mathcal{D}}$ are full-dimensional polytopes such that the vertices form a linear code. In the following, we classify all polytopes with this property. For a convex polytope P, let $V(P)$ denote the vertex set of P. For $n \in \mathbb{N}$ we denote $C_n := [0,1]^n$, unit cube and $W_n := \{0,1\}^n = V(C_n)$. its vertices. Hence $(W_n, \oplus)$ is an Abelian group that is canonically isomorphic to $(\mathbb{F}_2^n, \oplus)$. We consider W_n as a subset of $\mathbb{R}^n$ and write "$\oplus$" whenever we mean addition modulo 2, while "+" means ordinary addition in $\mathbb{R}^n$.

In the following, we develop an algorithm that determines inductively all polytopes $P \subseteq \mathbb{R}^n$ with $V(P) \subseteq W_n$ satisfying the following conditions:

(I) $(V(P), \oplus)$ is a subgroup of $(W_n, \oplus)$.
(II) P has dimension n.

The number of vertices of such a polytope is a power of two. Of course, C_n satisfies (I) and (II), but there are no further such polytopes when $n = 1$ and $n = 2$. For $n = 3$, the 3-dimensional regular simplex S with

$$V(S) = \{(0,0,0), (1,1,0), (1,0,1), (0,1,1)\} \tag{2.82}$$

satisfies (I) and (II), too. More generally, by [Wen06, Theorem 2.2], we have

Proposition 2.4.8. *For $n \geq 3$, the following statements are equivalent:*

(1) $(W_n, \oplus)$ contains some subgroup U such that $\operatorname{conv}(U)$ is a regular simplex of dimension n.
(2) $n + 1$ is some power of 2.

In the case $n = 3$, the 3-cube, as well as the regular simplex mentioned above, are the only polytopes satisfying conditions (I) and (II). Note that also

$$\{(0,0,0), (1,0,0), (0,1,1), (1,1,1)\} \tag{2.83}$$

determines a subgroup of $(W_3, \oplus)$; however, its convex hull has dimension two. For fixed $n \geq 2$, define the bijections $\pi_0 : \mathbb{R}^n \times \{0\} \to \mathbb{R}^n$ and $\pi_1 : \mathbb{R}^n \times \{1\} \to \mathbb{R}^n$ by

$$\begin{aligned} \pi_0(x_1, \ldots, x_n, 0) &:= (x_1, \ldots, x_n), \\ \pi_1(x_1, \ldots, x_n, 1) &:= (x_1, \ldots, x_n). \end{aligned} \tag{2.84}$$

For $1 \leq i \leq n$ put

$$\begin{aligned} H_i &:= \{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_i = 0\}, \\ H_i' &:= \left\{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_i = \frac{1}{2}\right\}. \end{aligned} \tag{2.85}$$

Moreover, let $z := \left(\frac{1}{2}, \ldots, \frac{1}{2}\right)$ denote the center of C_n. To determine recursively all 0/1-polytopes $P \subseteq \mathbb{R}^n$ that fulfill (I) and (II), we prove

Proposition 2.4.9. *Suppose that $n \geq 2$ and that $P \subseteq \mathbb{R}^n$ is a 0/1-polytope satisfying (I) and (II). Assume that $(U, \oplus)$ is a subgroup of $(V, \oplus) := (V(P), \oplus)$ with $|V : U| = 2$. Then the following statements are equivalent:*

(i) The polytope $Q \subseteq \mathbb{R}^{n+1}$, given by

$$Q = \operatorname{conv}(\pi_0^{-1}(U) \cup \pi_1^{-1}(V \setminus U)), \tag{2.86}$$

has dimension $n + 1$.
*(ii) There **does not** exist an index i with $1 \leq i \leq n$ such that $U \subseteq H_i$. In other words, none of the affine hyperplanes H_i' separates $\operatorname{conv}(U)$ and $\operatorname{conv}(V \setminus U)$.*
(iii) One has $z \in \operatorname{conv}(U) \cap \operatorname{conv}(V \setminus U)$.
(iv) One has $\operatorname{conv}(U) \cap \operatorname{conv}(V \setminus U) \neq \emptyset$.

Proof. **(i) ⇒ (ii):** Suppose that $U \subseteq H_i$ holds for some i with $1 \leq i \leq n$. Put

$$\tilde{H} := \left\{(x_1, \ldots, x_n, x_{n+1}) \in \mathbb{R}^{n+1} : x_{n+1} = x_i\right\}. \tag{2.87}$$

Since $P = \operatorname{conv}(V)$ has dimension n and since $|V : U| = 2$, we must have $x_i = 1$ whenever $(x_1, \ldots, x_n) \in V \setminus U$. This means that $V(Q)$, and hence also Q, is contained

in the n-dimensional hyperplane $\tilde{H}$, in contradiction to (i).
(ii) $\Rightarrow$ (iii): For $1 \leq i \leq n$, let $\alpha_i : \mathbb{F}_2^n \to \mathbb{F}_2$ denote the linear map given by $\alpha_i(x_1 \dots, x_n) := x_i$. By assumption, $\alpha_{i\restriction U}$ is surjective for $1 \leq i \leq n$. Hence we have

$$|\{u \in U : \alpha_i(u) = 0\}| = |\{u \in U : \alpha_i(u) = 1\}| \quad \text{for } 1 \leq i \leq n. \tag{2.88}$$

This means that

$$z = \frac{1}{|U|} \sum_{u \in U} u \in \operatorname{conv}(U). \tag{2.89}$$

Now fix $v_1 \in V \setminus U$, since $V \setminus U = \{v_1 \oplus u : u \in U\}$, we get also

$$|\{v \in V \setminus U : \alpha_i(v) = 0\}| = |\{v \in V \setminus U : \alpha_i(v) = 1\}| \quad \text{for } 1 \leq i \leq n, \tag{2.90}$$

and hence

$$z = \frac{1}{|V \setminus U|} \sum_{v \in V \setminus U} v \in \operatorname{conv}(V \setminus U). \tag{2.91}$$

(iii) $\Rightarrow$ (iv) is trivial.

(iv) $\Rightarrow$ (i): Consider the projection $\pi : \mathbb{R}^{n+1} \to \mathbb{R}^n$ given by

$$\pi(x_1, \dots, x_n, x_{n+1}) = (x_1, \dots, x_n). \tag{2.92}$$

Suppose that the assertion is wrong; hence Q is contained in some n-dimensional homogeneous hyperplane $G \subseteq \mathbb{R}^{n+1}$. Since

$$\pi(V(Q)) = U \dot{\cup} (V \setminus U) = V, \tag{2.93}$$

the polytope Q has the same dimension as $P = \operatorname{conv}(V)$, that is n. Thus, the restriction $\pi_{\restriction G}$ is a linear isomorphism from G onto $\mathbb{R}^n$, and there exists some $\mathbb{R}$-linear map $\alpha : \mathbb{R}^n \to G$ satisfying

$$(\alpha \circ \pi)(w) = w \qquad \text{for all } w \in G. \tag{2.94}$$

By definition of $V(Q)$, this means:

$$\begin{aligned} \alpha(U) &= \pi_0^{-1}(U), \\ \alpha(V \setminus U) &= \pi_1^{-1}(V \setminus U). \end{aligned} \tag{2.95}$$

Hence, $\alpha(\operatorname{conv}(U)) = \operatorname{conv}(\alpha(U))$ and $\alpha(\operatorname{conv}(V \setminus U)) = \operatorname{conv}(\alpha(V \setminus U))$ are linearly separated by the affine hyperplane

$$K := \left\{ (x_1, \dots, x_n, x_{n+1}) \in \mathbb{R}^{n+1} : x_{n+1} = \frac{1}{2} \right\}. \tag{2.96}$$

By (iv) this is impossible. □

EXAMPLE 2.4.10. We investigate the statement of Proposition 2.4.9 for polytopes corresponding to an arbitrary collection $\mathcal{D}$. Consider the matrix that has as its rows the vectors $f_{\mathcal{D}}(x)$, where $\mathcal{D} = 2^{[N]} \setminus \{\emptyset\}$. The rows are labeled by the binary strings of length N, that is by $\mathcal{X}$, while the columns are indexed by the nonempty subsets of $[N]$. Therefore the rows of this matrix are the coordinates of the vertices of a simplex:

x	$\{1\}$	$\cdots$	$\{N\}$	$\{1,2\}$	$\cdots$	$\{1,2,3\}$	$\cdots$	$[N]$
$(00\dots0)$	0	$\cdots$	0	0	$\cdots$	0	$\cdots$	0
$(00\dots1)$	0	$\cdots$	1	$f_{\{1,2\}}(x)$	$\cdots$	$f_{\{1,2,3\}}(x)$	$\cdots$	$f_{[N]}(x)$
$\vdots$	$\vdots$	$\dots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$
$(11\dots1)$	1	$\cdots$	1	$f_{\{1,2\}}(x)$	$\cdots$	$f_{\{1,2,3\}}(x)$	$\cdots$	$f_{[N]}(x)$

We note the following facts:

- The columns of this matrix are indexed by the $2^N - 1$ nonzero, binary strings of length N.
- There are $2^N - 1$ subgroups U of index 2 of the N-cube, which correspond to the columns of the matrix. To define them let a column with label B, say, be fixed and put $U := \{x \in \mathcal{X} : f_B(x) = 0\}$. The maps $f_B : \mathcal{X} \to \{0, 1\}$ are exactly the $2^N - 1$ surjective homomorphisms having the nontrivial subgroups as their kernels.
- The vertices of every polytope $\mathcal{F}_{\mathcal{D}}$ are given by the rows of this matrix after deleting columns that correspond to sets not in $\mathcal{D}$.
- In particular, by restriction to the first N columns, we get the vertices of the N-cube, corresponding to the binary independence model.

Now, assume that P is the N-cube. We choose a column of the matrix, corresponding to a subgroup of index 2. There are two possibilities. If we choose a column corresponding to an atom, then *(ii)* is wrong, the dimension does not grow when adding this column to the coordinates (as we have doubled a coordinate). If, on the other hand, we choose a column corresponding to a set B with cardinality two or more, then we are in the situation of Proposition 2.4.9, since *(ii)* holds. The lift (2.86) is full-dimensional, and its vertices are given by the submatrix with columns $\{1\}, \ldots, \{N\}, B$. Continuing from here, choosing another subgroup, the dimension grows if and only if it does not correspond to one of the sets $\{1\}, \ldots, \{N\}, B$. Iteratively, the choices narrow down, and finally, when all columns have been chosen, the polytope Q is a simplex.

We formalize this procedure now. For fixed polytope P as in Proposition 2.4.9, put

$$U_i := V(P) \cap H_i \qquad \text{for } 1 \le i \le n. \tag{2.97}$$

Conditions (I) and (II) imply that $|V(P) : U_i| = 2$, whenever $1 \le i \le n$. Based on the equivalence of (i) and (ii) in Proposition 2.4.9, we are now able to prove that the following algorithm yields recursively all 0/1-polytopes satisfying (I) and (II).

ALGORITHM 2.4.11.
Initialization for $n = 1$***:***

- $\mathfrak{P}_1 := \{[0, 1]\}$.

Step $n \to n + 1$***:*** *Based on* $\mathfrak{P}_n$ *construct a new set* $\mathfrak{P}_{n+1}$ *consisting of all polytopes* Q *such that there exists* $P \in \mathfrak{P}_n$ *with*

- $Q = P \times [0, 1]$ *or*
- $Q \subseteq \mathbb{R}^{n+1}$ *with*

$$V(Q) = \pi_0^{-1}(U) \cup \pi_1^{-1}(V(P) \setminus U), \tag{2.98}$$

where U *runs through all subgroups of* $(V(P), \oplus)$ *with* $|V(P) : U| = 2$ *and* $U \neq U_i$ *for* $1 \le i \le n$.

In the case $Q = P \times [0, 1]$, the number of vertices is doubled, while in the other cases the number of vertices of Q equals the number of vertices of P. The two possible operations commute in the following sense. Starting from some cube W_n, lifting it to W_{n+1} and then choosing a subgroup U to apply the lift (2.98) gives the same polytope as choosing the subgroup $\pi(U)$ from W_n and then taking the prism over the lifted polytope, where $\pi : \mathbb{R}^{n+1} \to \mathbb{R}^n$ is the canonical projection. Therefore, all polytopes that are constructed by the algorithm can be thought of as lifted cubes W_n. The classification is complete with:

THEOREM 2.4.12. *For all* $n \in \mathbb{N}$*, the set* $\mathfrak{P}_n$ *in Algorithm 2.4.11 consists of all* n*-dimensional 0/1-polytopes that satisfy conditions (I) and (II).*

Proof. First we show that all polytopes $Q \in \mathfrak{P}_{n+1}$ satisfy conditions (I) and (II), with n replaced by $n+1$. This is clear in the case of the prism $Q = P \times [0,1]$. If Q satisfies (2.98), then clearly $(V(Q), \oplus)$ is a subgroup of $(W_{n+1}, \oplus)$, because U is a subgroup of $(V(P), \oplus)$ with $|V(P) : U| = 2$. Moreover, (ii) $\Rightarrow$ (i) in Proposition 2.4.9 implies that Q has dimension $n+1$, because $U \neq U_i$ for $1 \le i \le n$. Hence, Q satisfies conditions (I) and (II).

Vice versa, assume that $Q \subseteq \mathbb{R}^{n+1}$ fulfills (I) and (II). Consider again the projection $\pi : \mathbb{R}^{n+1} \to \mathbb{R}^n$ onto the first n coordinates, and put $P := \pi(Q)$. Since Q has dimension $n+1$, P has dimension n. If $\pi_{\restriction V(Q)}$ is not injective, then Q is the prism $P \times [0,1]$, because $(V(Q), \oplus)$ is a subgroup of $(W_{n+1}, \oplus)$. If $\pi_{\restriction V(Q)}$ is injective, put

$$U := \{(x_1, \dots, x_n) \in V(P) | (x_1, \dots, x_n, 0) \in Q\}. \tag{2.99}$$

Then $(U, \oplus)$ is a subgroup of $(V(P), \oplus)$ with $|V(P) : U| = 2$, because Q has dimension $n+1$. Moreover, equation (2.98) holds for U as just defined. Finally, Proposition 2.4.9, (i) $\Rightarrow$ (ii) shows that $U \neq U_i$ for $1 \le i \le n$. Hence, our algorithm includes the determination of Q. □

As a first application of Theorem 2.4.12 we can count the number of n-dimensional polytopes that satisfy conditions (I) and (II). Let $c_n := |\mathfrak{P}_n|$. For $1 \le k \le n$, let $c_n(k)$ denote the number of all 0/1-polytopes $P \subseteq \mathbb{R}^n$ with $|V(P)| = 2^k$ that satisfy (I) and (II). Then one has

$$c_n = \sum_{k=1}^{n} c_n(k). \tag{2.100}$$

We have $c_n(k) = 0$ for $2^k \le n$, because a polytope with at most n vertices cannot have dimension n. Clearly $c_n(n) = 1$ for all $n \in \mathbb{N}$. As mentioned already in Example 2.4.10, a 0/1-polytope that satisfies (I), (II), and $|V(P)| = 2^k$, has among its vertices exactly $2^k - 1$ subgroups of index 2. Hence by ignoring the groups $U_i = V(P) \cap H_i$ for $1 \le i \le n$, we get

Corollary. *For $k \le n < 2^k$ one has*

$$c_{n+1}(k) = c_n(k-1) + c_n(k)(2^k - n - 1). \tag{2.101}$$

The first few values for $c_n(k)$ are given in the Table 1. It is easy to compute this

n \ k	1	2	3	4	5	6	7	8	c_n
1	1								1
2	0	1							1
3	0	1	1						2
4	0	0	5	1					6
5	0	0	15	16	1				32
6	0	0	30	175	42	1			248
7	0	0	30	1605	1225	99	1		2960
8	0	0	0	12870	31005	6769	219	1	50864

Table 1. The number of n-dimensional 0/1-polytopes with 2^k vertices that form a group.

number also for larger values of n. For instance

$$\begin{gathered} c_{28} = 718897730072178204358180468879825453986397667929112558174208, \\ c_{100} \approx 2.77 \cdot 10^{644}. \end{gathered} \tag{2.102}$$

Using the Corollary we can show that among the full-dimensional 0/1-polytopes with 2^k vertices the convex hulls of linear codes are exceptional. For $1 \leq k \leq n$, let $d_n(k)$ denote the number of all 0/1-polytopes with 2^k vertices satisfying only condition (II). Hence d_n, the number of all 0/1-polytopes of dimension n trivially satisfies

$$d_n \geq \sum_{k=1}^{n} d_n(k). \tag{2.103}$$

Moreover, we get

PROPOSITION 2.4.13. *(i) For $4 \leq n < 2^k < 2^n$, one has*

$$d_n(k) \geq 2^k(2^n - 2^k)c_n(k) > n2^{n-1}c_n(k). \tag{2.104}$$

(ii) We have

$$\lim_{n\to\infty} \frac{c_n}{d_n} = 0. \tag{2.105}$$

PROOF. **(i):** Suppose that U is a proper subgroup of $(W_n, \oplus)$ with $\dim(\mathrm{conv}(U)) = n$ and $|U| = 2^k$. If U' is another subgroup of $(W_n, \oplus)$ with $|U'| = |U|$, then we have

$$\left|U \cap U'\right| \leq 2^{k-1} < 2^{n-1}, \tag{2.106}$$

and hence,

$$\left|U \setminus U'\right| \geq 2^{k-1} > \frac{n}{2} \geq 2. \tag{2.107}$$

This means

$$\left|U \setminus U'\right| \geq 3. \tag{2.108}$$

There are $2^k(2^n - 2^k)$ subsets V of W_n with $|V| = 2^k$ and $|V \setminus U| = |U \setminus V| = 1$; namely, these are all sets of the form

$$V = (U \setminus \{u_0\}) \cup \{v_0\} \qquad \text{with } u_0 \in U, v_0 \in W_n \setminus U. \tag{2.109}$$

For V as in (2.109), we get $\dim(\mathrm{conv}(V)) = n$, because otherwise, $U \setminus \{u_0\}$ would be contained in a unique hyperplane H with $v_0 \in H$, a contradiction to $v_0 \notin U$. Together with (2.108), we obtain the first inequality in (2.104). The second one is trivial in view of $2^k \leq 2^{n-1}$.

(ii): By (2.100), (2.103), and (2.104) we get for $n \geq 4$:

$$\begin{aligned} \frac{c_n}{d_n} &\leq 2\frac{c_n - 1}{d_n - 1} \\ &\leq 2\left(\sum_{k=1}^{n-1} c_n(k)\right)\left(\sum_{k=1}^{n-1} d_n(k)\right)^{-1} \\ &\leq 2(n2^{n-1})^{-1} = \frac{2^{2-n}}{n}. \end{aligned} \tag{2.110}$$

This proves the second statement. □

2.5. Support Sets and Neighborliness

In this final section on marginal polytopes we will deduce another corollary of Theorem 2.3.5, this time as a statement about marginal polytopes and support sets of hierarchical models. Before that we show an explicit example where knowledge about the marginal polytope allows determination of the complete face lattice, and thus all support sets of the corresponding hierarchical model.

EXAMPLE 2.5.1 (Supports in the binary no-n-way interaction model). We consider the binary hierarchical model whose simplicial complex is the boundary of an n-simplex. If $n = 3$, this model is called the no-3-way interaction model and its Markov bases have been recognized to be arbitrarily complicated [LO06], so we cannot hope to find an easy description of the oriented circuits. However, if we restrict ourselves to binary random variables $x = (x_i)_{i=1}^n \in \mathcal{X} := \{0,1\}^n$, the structure is very simple. In this case the exponential family is of dimension $2^n - 2$, i.e. of codimension 1 in the simplex, so $\ker A$ is one-dimensional. It is spanned by the "parity function":

$$e_{[n]}(x) := \begin{cases} -1 & \text{if } \sum_{i=1}^n x_i \text{ is odd,} \\ 1 & \text{otherwise.} \end{cases} \tag{2.111}$$

This vector is also the (up to sign) unique elementary circuit of $\ker A$. Using Theorem 1.3.13 we can describe the face lattice of the marginal polytope (i.e. convex support) $P^{(n-1)}$: A set $\mathcal{Y} \subsetneq \{0,1\}^n$ is a support set if and only if it does not contain all configurations with even parity, or all configurations with odd parity. Recall that a polytope is called k-neighborly if the convex hull of any k or less vertices is a face of the polytope. It can be seen that no d-polytope apart from the simplex is more than $\lfloor \frac{d}{2} \rfloor$-neighborly. Therefore a polytope, achieving this bound is simply called a *neighborly polytope*. It follows that $P^{(n-1)}$ is *neighborly*, since no set of cardinality less than 2^{n-1} can contain all configurations with even or odd parity. We can easily count the support sets by counting the nonfaces of the corresponding marginal polytope, i.e. all sets $\mathcal{Y}$ that contain either the configurations with even parity, or the configurations with odd parity. Let s_k be the number of support sets of cardinality of k, i.e. the number of faces with k vertices. It is given by:

$$s_k = \binom{2^n}{k} - 2\binom{2^{n-1}}{k - 2^{n-1}}, \tag{2.112}$$

where we adopt the convention that $\binom{m}{l} = 0$ if $l < 0$. Since this polytope has only one affine dependency (2.111), which includes all the vertices, we see that it is *simplicial*, i.e. all its faces are simplices. It follows that f_k, the components of the f-vector, are given by $f_k = s_{k-1}$. Neighborliness of marginal polytopes is discussed in Section 2.5.1 and is the topic of [Kah10b].

Altogether we have determined the face lattice of the polytope, which means that we know the "combinatorial type" of the polytope. It turns out that the face lattice of $P^{(n-1)}$ is isomorphic to the face lattice of the $(2^n - 2)$-dimensional *cyclic polytope* with 2^n vertices.

2.5.1. Neighborliness of Marginal Polytopes. The following is a main result of this thesis.

THEOREM 2.5.2. *Let g be the minimal cardinality among the nonfaces of Δ.*

Geometric Formulation: *The marginal polytope is $2^{g-1} - 1$ neighborly.*

Probabilistic Formulation: *Every distribution p with $|\mathrm{supp}(p)| < 2^{g-1}$ is contained in $\overline{\mathcal{E}_\Delta}$.*

PROOF. The probabilistic formulation is easy to see. Just observe that by Theorem 2.3.5 each monomial appearing in the set of generators $\left\{p^{m^+} - p^{m^-} : m \in M\right\}$ has cardinality of its support bounded from below by 2^{g-1}. Therefore a p with $|\mathrm{supp}(p)| < 2^{g-1}$ must fulfill the defining equations trivially. The geometric formulation is immediate by Theorem 1.2.14. $\square$

Remark 2.5.3 (The bound is sharp). On first sight one would maybe expect a better neighborliness property in the nonbinary cases, for instance if every random variable is ternary. However, one can see that the bound is sharp. Already for the "no-three-way-interaction" model with ternary random variables, given by $N = \{1,2,3\}, \mathcal{X}_i = \{0,1,2\}$ for $i = 1,2,3$ and $\Delta = \{B \subseteq \{1,2,3\} : |B| \le 2\}$, one has square-free generators of degree 4. They can be computed with 4ti2 [4ti207] or looked up in the MBDB, as described in Section 2.3.3. Then a p that is nonzero exactly on the positive support is a counterexample for any improvement of Theorem 2.5.2.

Remark 2.5.4 (Maximizing multi-information). The *multi-information* is an entropic quantity that generalizes mutual information to more than two random variables. Denoting $H(p) := -\sum_{x\in\mathcal{X}} p(x)\log p(x)$ the entropy of p, and $H_i(p) := -\sum_{x\in\mathcal{X}_i} p_{\{i\}}(x)\log p_{\{i\}}(x)$ the marginal entropy for $i \in N$, it is defined as

$$\mathcal{M}(p) := \sum_{i\in N} H_i(p) - H(p). \tag{2.113}$$

It can be seen that this function is exactly the Kullback-Leibler divergence $D(p \parallel \mathcal{E}_1) := \min_{q\in\overline{\mathcal{E}_1}} D(p \parallel q)$ of p from the family $\mathcal{E}_1$. An interesting problem, considered in [Ay02; AK06], is to maximize the multi-information, or more generally, the divergence from a given exponential family. In [AK06] all global maximizers in the binary case are classified giving their support sets, and the question to construct a low-dimensional family containing all maximizers is raised. It holds that any global maximizer is supported on two elements only. For a general exponential family $\mathcal{E}$, in Proposition 3.2 of [Ay02] it is shown that any p^* whose maximum likelihood estimate lies in the open family $\mathcal{E}$ and which *locally* maximizes $D(p \parallel \mathcal{E})$ satisfies

$$|\mathrm{supp}(p^*)| \le \dim(\mathcal{E}) + 1. \tag{2.114}$$

In [Mat07] it was shown that the first condition, that p^* has a maximum likelihood estimate with full support, is not needed. Now, let $\Delta_k := \{B \subseteq N : |B| \le k\}$ denote the uniform simplicial complex of order k, then we can deduce a minimal k^*, such that $\overline{\mathcal{E}_{\Delta_k}}$ contains all local maximizers. Construction of such low-dimensional families can be interesting in the modeling of cognitive systems using information theory. By Theorem 2.5.2 we find

Corollary. *Any local maximizer of $\mathcal{M}$ is contained in the closure of the uniform hierarchical model of order $k^* \ge \log_2(\dim(\mathcal{E}) + 2)$.*

Note that for the multi-information in the binary case we have $\dim(\mathcal{E}_1) = n$. The problem of understanding local and global maximizers of information divergence is not solved, but recently there has been significant progress, see [Mat07; Rau09].

CHAPTER 3

Primary Decomposition of Binomial Ideals

3.1. Solutions of Polynomial Equations

Hierarchical models are defined by binomial prime ideals. A consequence of this is that they admit nice parameterizations and can be seen as manifolds with boundary. Not every statistical model has this property. Here we discuss algebraic techniques to treat the support set problem for models that are initially defined by binomial equations, such as (saturated) conditional independence (CI) models. A typical phenomenon for such models is that dimension varies locally. The technique of primary decomposition allows to write a reducible variety as the union of irreducible ones. It does even more than this geometric operation. It decomposes the defining ideal in terms of primary ideals and thereby, geometrically, decomposes the scheme of I into its parts. For statistical interpretations the additional information that a scheme carries, compared to a variety, is usually not used. It is therefore natural to expect that ideals in algebraic statistics are radical ideals or can be replaced by their radicals. A recent conjecture in [ESU10] goes in this direction. There certain stochastic processes are defined through binomial ideals and it is conjectured that these ideals are radical. This would imply a minimality of description as no additional algebraic information is contained. The conjecture turns out to be false as we will see in Section 3.3.

The main content of this chapter is a description of the Macaulay 2 [GS; EGSS01] package *Binomials*, which is the first implementation of specialized algorithms for binomial ideals. This package achieves significant speed-ups over general purpose implementations. As a warm-up the reader is invited to review some basics of solving systems of polynomial equations in the following section.

3.1.1. Some Commutative Algebra. Linear Algebra gives a complete answer to the problem of solving systems of linear equations and describing the geometry of the set of solutions. Primary decomposition strives to achieve the same for polynomial equations of arbitrary degree. *Varieties*, the sets of solutions of polynomial systems, have a very sophisticated structure. The most prominent features, distinguishing them from smooth manifolds, are the possible presence of singularities and that the dimension is not a global invariant anymore, it is a local property. This is one of the reasons why it is not generally possible to parametrize the solutions of a polynomial system continuously.

We establish primary decomposition as an approximation to solving systems of polynomial equations. Consider a polynomial ring in n indeterminates: $S := \Bbbk[x] := \Bbbk[x_1, \dots, x_n]$, where $\Bbbk$ is a field. An ideal $I \subseteq S$ is an additive subgroup of S that is closed under multiplication with ring elements $IS \subseteq I$. Hilbert's Basis Theorem guarantees that every ideal in S is finitely generated, meaning that their exist finitely many polynomials $f_1, \dots, f_s$, such that

$$I = \langle f_1, \dots, f_s \rangle := \left\{ \sum_{i=1}^{s} g_i f_i,\ g_1, \dots, g_s \in S \right\}. \tag{3.1}$$

The polynomials f_i need not be unique and not even their number is constant over different generating systems. In a certain sense nice systems of generators are offered by Gröbner bases [CLO96]. The *variety* of I, denoted $V(I)$, is the set of common zeros of I.

$$V(I) := \{x \in \Bbbk^n : f(x) = 0, \forall f \in I\}. \tag{3.2}$$

Conversely, given a set $V \subseteq \Bbbk^n$, we define

$$I(V) := \{f \in S : f(x) = 0, \forall x \in V\} \tag{3.3}$$

the *ideal* of V. It is easy to see that $I(V)$ is in fact an ideal. A variety in $\Bbbk^n$ is called *irreducible* if it is not the union of two proper subvarieties.

Example 3.1.1. Consider, for fixed $p, q \in \mathbb{C}$, the principal ideal

$$I = \langle x^2 + px + q \rangle \subseteq \mathbb{C}[x]. \tag{3.4}$$

Its variety consists of the two points $V(I) = \left\{-\frac{p}{2} \pm \sqrt{\frac{p^2}{4} - q}\right\}$. As each point is the solution of a linear equation this variety is reducible. It is instantly visible that the definitions of variety and reducibility depend on the field.

It is often convenient to describe varieties using the ring of polynomial functions defined on them, the *coordinate ring*. It is given by the quotient (of Abelian groups) S/I. Properties of ideals often give dual properties of the coordinate ring and vice versa. S/I has the structure of an S-algebra, in particular it is an S-module [Hun74]. The geometry of the variety can be studied in terms of properties of S/I. A *zerodivisor* in any commutative ring R is an element $r \in R$ such that there exists a nonzero $s \in R$ with $rs = 0$. For radical ideals, zerodivisors in S/I indicate that $V(I)$ is not irreducible. This means that we find two nonzero polynomial functions such that their product is zero. This can happen if the variety consists of two “parts” and each polynomial is nonzero only on one of the parts. In the case of a general ideal, its variety is irreducible if each zerodivisor f is *nilpotent*, that is, a power of f vanishes. It is visible here that ideals encode more information than just the variety. The geometric object described by an ideal in a commutative ring (with unity) is that of a *scheme* [EH01].

The algebraic way to remove zerodivisors in S/I is to enlarge I and thereby “putting the zerodivisors to zero”. The geometric intuition here is to remove a part of the variety where the zerodivisor takes nonzero values and thus approximate irreducible parts of $V(I)$. For an ideal $I \subseteq S$, and $f \in S/I$, we define the ideal $(I : f) := \{g \in S : fg \in I\}$. Again, we have to take care of the case that I is not radical. In this case f might still be a zerodivisor. We define $(I : f^\infty) := \{g \in S : f^n g \in I, n \in \mathbb{N}\}$. This definition makes sense since S is Noetherian and the chain of ideals

$$I \subseteq (I : f) \subseteq ((I : f) : f) \subseteq (((I : f) : f) : f) \subseteq \dots \tag{3.5}$$

becomes stationary at some point. It is visible that f is a nonzerodivisor if and only if $(I : f) = I$. Let us define ideals with certain restrictions on zerodivisors in the respective coordinate rings. An ideal I is *maximal* if S/I is a field. An ideal P is *prime* if S/P has no zerodivisors: $(P : f) = P$ for all $f \in S/P$. It is called *primary* if in S/P every zerodivisor is nilpotent. In Definition 3.2.2 we will define cellular ideals, which also fit into this hierarchy.

The varieties of prime and primary ideals are irreducible. We say that a prime P is *associated* to I if there exists $f \in S$ such that $(I : f) = P$. The intuition here is that the irreducible variety of P forms part of the variety of I. It can be seen that each ideal has a finite nonempty set of associated primes, denoted $\mathrm{Ass}(S/I)$. The associated primes can be contained in each other and thus form a poset. The minimal elements are called *minimal primes*, the remaining elements are called *embedded* as their varieties are contained in the varieties of the minimal primes. The *radical* of I is the intersection of its minimal primes $\operatorname{Rad}(I) := \bigcap \{P : P \text{ minimal over } I\}$. The radical of a primary ideal Q is a prime ideal $P = \operatorname{Rad}(Q)$ with the same variety. The prime contains only the geometric information about the variety, forgetting about the scheme of Q. A foundational theorem states that, over algebraically closed fields radical ideals and varieties are in 1:1 correspondence:

THEOREM 3.1.2 (Hilbert's Nullstellensatz). *In polynomial rings over algebraically closed fields, for any ideal I:*

$$I(V(I)) = \operatorname{Rad}(I) \tag{3.6}$$

The main result for analyzing solutions of polynomial systems is:

THEOREM 3.1.3. *Every ideal $I \subseteq S$ is the intersection of finitely many primary ideals,*

$$I = Q_1 \cap Q_2 \cap \ldots \cap Q_r, \tag{3.7}$$

where the primes $P_i = \operatorname{Rad}(Q_i)$ are distinct, unique, and associated to I. The primary ideals corresponding to minimal primes are unique.

EXAMPLE 3.1.4. We consider polynomial systems given through conditional independence statements. Recall Definition 2.2.13 where we defined equations given by CI-statements. A *CI-ideal* is an ideal defined through equations of the form (2.43). Let $\mathcal{C} = \{\, A_i \perp\!\!\!\perp B_i \,|C_i \; : i \in [r]\}$ be a collection of CI-statements, each defining a prime ideal $I_{A_i \perp\!\!\!\perp B_i \,|C_i}$ in the polynomial ring $\Bbbk[p_x : x \in \mathcal{X}]$. The CI-ideal associated with the collection is the sum of these ideals

$$I_{\mathcal{C}} := I_{A_1 \perp\!\!\!\perp B_1 \,|C_1} + \cdots + I_{A_r \perp\!\!\!\perp B_r \,|C_r} . \tag{3.8}$$

A longer discussion on the subject can be found in [DSS09; Stu02].

Consider three binary random variables X_1, X_2, X_3. The polynomial ring is given by $\Bbbk[p_{111}, p_{112}, p_{121}, p_{122}, p_{211}, p_{212}, p_{221}, p_{222}]$. The conditional independence ideal of the statement $X_1 \perp\!\!\!\perp X_2 \,|X_3$ is the binomial ideal

$$I_{X_0 \perp\!\!\!\perp X_1 \,|X_2} = \langle p_{111}p_{221} - p_{121}p_{211}, p_{112}p_{222} - p_{122}p_{212} \rangle . \tag{3.9}$$

In contrast to that, the independence $X_1 \perp\!\!\!\perp X_2$ is given by the principal ideal

$$I_{X_0 \perp\!\!\!\perp X_1} = \langle (p_{111} + p_{112})(p_{221} + p_{222}) - (p_{211} + p_{212})(p_{121} + p_{122}) \rangle . \tag{3.10}$$

A typical task, solvable by primary decomposition, is to describe, interpret, and parametrize the sets of solutions of these equations. Thus CI-ideals are a natural place to apply primary decomposition, and first results, supported by computations with *Binomials*, have already emerged [HHH+10; Fin09].

3.1.2. Support Sets and Primary Decomposition. Let I be a binomial ideal, for instance the CI-ideal in (3.9). The varieties of binomial ideals have special features, each of them posses a prime component given by the toric ideal $(I : (\prod_{i=1}^n x_i)^\infty)$. This is the only component whose variety can contain strictly positive probability distributions. For studying its support sets, the theory of Chapter 1 applies. Any other primary component contains monomials p^m, and thus its variety is a strict subset of the union of the coordinate planes in $\mathbb{C}^n$. Probability distributions in these components must have zeros. After having restricted to a subset of variables of the polynomial ring, the situation is essentially the same as for the toric component. This is one of the main results of [ES96], each associated prime of a binomial ideal, after a change of coordinates, is a toric ideal. For this reason we are facing two layers of the support set problem when dealing with statistical models defined through not necessarily prime binomial ideals. Apart from studying the geometry of the individual components, the problem considered in Chapters 1 and 2, one has to decompose the variety into its components first. The second problem is discussed now.

3.2. Specialized Algorithms for Binomial Ideals

3.2.1. Binomial Ideals. In the following we present *Binomials*, a software package which provides specialized algorithms for binomial ideals, allowing for a significant speed-up of common computations like primary decomposition. Central parts of the implemented algorithms go back to Eisenbud and Sturmfels' paper [ES96], which develops the theory of binomial ideals in depth. Surprisingly their algorithms had never been implemented, but the software described here filled the gap. To demonstrate the power of our approach we show how *Binomials* was used to compute primary decompositions of commuting birth and death ideals of [ESU10], yielding a counterexample for a conjecture therein. The material here is also the content of [Kah10a].

A monomial ideal is an ideal generated by monomials, a binomial ideal is one whose generators can be chosen as binomials. A *pure difference binomial ideal* is an ideal whose generators are differences of monic monomials. For monomial ideals, central concepts like Gröbner bases or primary decompositions can be defined directly on the exponent vectors of the monomials generating the ideal. In this sense the whole theory is very combinatorial. For binomial ideals the situation is more complicated, but it can be turned combinatorial too. Note that any ideal can be written generated by "trinomials" if one allows additional variables.

Binomial ideals occur in many applications. We have seen toric ideals, which are binomial prime ideals. They are exactly the defining ideals of toric varieties, given through their classical definition in [Ful93]. In the previous chapters we have seen the central role that they play in the description of exponential families. Finitely generated, affine, commutative semigroup rings are quotients of polynomial rings by pure difference binomial ideals [MS05, Chapter 7], and in Section 3.3 we will define commuting birth and death processes, which present an application of binomial ideals in probability theory.

Consider the polynomial ring $S = \Bbbk[x_1, \ldots, x_n]$ over a field $\Bbbk$ of characteristic zero. In the following, choices for $\Bbbk$ are the rationals $\mathbb{Q}$, their cyclotomic extensions $\mathbb{Q}(\xi_l)$, or the complex numbers $\mathbb{C}$. Primary decompositions of binomial ideals are not necessarily binomial as is easily seen on the ideal $\langle x^3 - 1\rangle$, which over $\mathbb{Q}$ decomposes as $\langle x - 1\rangle \cap \langle x^2 + x + 1\rangle$. However, if $\Bbbk$ is algebraically closed, binomial primary decompositions exist. The notions here are a bit involved. When speaking of primary decompositions, in the following we always mean primary decomposition into binomial ideals, and we have to extend the coefficient field where needed. For our implementation of primary decomposition we have restricted to pure difference binomial ideals, where cyclotomic extensions of $\mathbb{Q}$ suffice. In many applications it suffices to study this case. Examples include the semi-graphoid ideal [HMS$^+$08], conditional independence ideals [DSS09; Fin09; HHH$^+$10] and defining ideals of commutative semigroup rings [Gil84]. Throughout we use notation that tries to coincide with that in [ES96].

If one wants to avoid extending the coefficient field, or even stay in the class of pure difference binomial ideals, coarser decompositions are interesting. We discuss cellular decompositions in Section 3.2.2 and give a fast algorithm for computing the minimal primes of a binomial ideal in Section 3.2.4. In Section 3.2.3 we give an algorithm for finding the solutions of zero-dimensional pure difference binomial ideals and apply it to saturation of partial characters. This yields the primary decomposition in Section 3.2.5. Finally in Section 3.3 we study primary decompositions of commuting birth and death ideals and give a counterexample to Conjectures 5.3 and 5.9 in [ESU10].

In keeping with the introductory nature of this work, each section contains examples of how to do the discussed computations with the help of *Binomials*. These examples are thought of as a motivation and do not cover all of the functionality that is implemented. They are produced with version 0.5.4 of *Binomials*. The reader is encouraged to download the package, use it, and report experiences to the author. An online help is integrated. The first example shows how to get started.

EXAMPLE 3.2.1 (Downloading and using *Binomials*). *Binomials* and an auxiliary package for cyclotomic fields, called *Cyclotomic*, are separately available under the URL:

(3.11) `http://personal-homepages.mis.mpg.de/kahle/bpd/`

It is recommended to install the latest version of Macaulay 2 before using *Binomials*. To get started, run Macaulay 2, then load the package with

```
i1 : load "Binomials.m2"
--loading configuration for package "Binomials" from file [...]
--loading configuration for package "FourTiTwo" from file [...]
```

For *Binomials* to be loaded, the additional packages *FourTiTwo* and *Cyclotomic* are needed. The first is included in Macaulay 2 as of version 1.2, while the latter can be obtained together with *Binomials*. The Macaulay 2 system reports that it loaded additional configuration files. To make the documentation available the package should be installed. This can be done via

```
i2 : installPackage ("Binomials", RemakeAllDocumentation=>true)
```

After running this, help can be accessed via

```
i3 : help "Binomials"
```

which shows an overview of the functionality of *Binomials*.

3.2.2. Cell Decompositions of Binomial Varieties. Our analysis of a binomial variety starts with the decomposition of $\Bbbk^n$ into the 2^n algebraic tori, interior to the coordinate planes. Each of the coordinate planes is defined by a subset $\mathcal{E} \subseteq \{1, \dots, n\}$ of the indeterminate's indices. We denote the algebraic torus corresponding to $\mathcal{E}$ by

$$(\Bbbk^*)^{\mathcal{E}} := \{(x_1, \dots, x_n) \in \Bbbk^n : x_i \neq 0, i \in \mathcal{E} \text{ and } x_j = 0, \forall j \notin \mathcal{E}\}. \tag{3.12}$$

Geometrically, for an ideal $I \subseteq \Bbbk[x_1, \dots, x_n]$, we study *cellular decompositions*. Their components are the intersections of primary components which have generic points in a given cell $(\Bbbk^*)^{\mathcal{E}}$. The central definition is

DEFINITION 3.2.2. A proper binomial ideal I in a polynomial ring R, is called *cellular*, if each variable of R is either a nonzerodivisor or nilpotent modulo I.

Primary ideals and lattice ideals are cellular. The following explicit representation of cellular ideals is only a reformulation of the definition but useful in many ways.

LEMMA 3.2.3. *An ideal $I \subsetneq R$ is cellular if and only if there exists a set $\mathcal{E} \subseteq \{1, \dots, n\}$ of variable indices of R such that*

1. $I = \left(I : (\prod_{i\in\mathcal{E}} x_i)^\infty\right)$,
2. *For every $i \notin \mathcal{E}$, there exists a nonnegative integer d_i such that the ideal $\left\langle x_i^{d_i} : i \notin \mathcal{E} \right\rangle$ is contained in I.*

We call the set $\mathcal{E}$ the *cell indices* and the variables $\{x_i : i \in \mathcal{E}\}$, which are exactly the nonzerodivisors modulo I, the *cell variables*. We denote by $M(\mathcal{E})$ the ideal generated by the noncell variables, i.e. the variables $\{x_i : i \notin \mathcal{E}\}$. For any vector $d = (d_i)_{i \notin \mathcal{E}}$ of natural numbers we denote $M(\mathcal{E})^d := \left\langle x_i^{d_i} : i \notin \mathcal{E} \right\rangle$. With this notation, another useful representation of cellular ideals is given by the following Lemma. In [ES96] the ideal on the right hand side of (3.13) is denoted $I_{\mathcal{E}}^{(d)}$.

LEMMA 3.2.4. *An ideal I is cellular if and only if there exist a set $\mathcal{E} \subseteq \{1, \ldots, n\}$ and an exponent vector d, such that*

$$I = \left(\left(I + M(\mathcal{E})^d \right) : \left(\prod_{i \in \mathcal{E}} x_i \right)^{\infty} \right). \tag{3.13}$$

Radicals of cellular binomial ideals have a nice combinatorial structure, defined by the set $\mathcal{E}$, and a partial character, which we introduce next. For this let $\emptyset \neq \mathcal{E} \subseteq \{1, \ldots, n\}$ be any nonempty subset of the indices of variables and define the shorthand $\Bbbk[\mathcal{E}] := \Bbbk[x_i : i \in \mathcal{E}]$.

DEFINITION 3.2.5. A *partial character* is a pair $(\mathcal{L}, \sigma)$, consisting of an integer lattice $\mathcal{L} \subseteq \mathbb{Z}^{\mathcal{E}}$ and a map $\sigma : \mathcal{L} \to \Bbbk^*$, that is a homomorphism from the additive group $\mathcal{L}$ to the multiplicative group $\Bbbk^*$. For each integer lattice $\mathcal{L} \subseteq \mathbb{Z}^{\mathcal{E}}$, we define its *saturation*

$$\mathrm{Sat}(\mathcal{L}) := \left\{ m \in \mathbb{Z}^{\mathcal{E}} : dm \in \mathcal{L} \text{ for some } d \in \mathbb{Z} \right\}. \tag{3.14}$$

A lattice $\mathcal{L} \subseteq \mathbb{Z}^{\mathcal{E}}$ is called *saturated* if it satisfies $\mathrm{Sat}(\mathcal{L}) = \mathcal{L}$. A partial character $(\mathcal{L}, \sigma)$ is called *saturated* if $\mathcal{L} = \mathrm{Sat}(\mathcal{L})$, and it is called a *saturation* of a partial character $(\mathcal{L}', \sigma')$, provided that $\mathcal{L} = \mathrm{Sat}(\mathcal{L}')$ and $\sigma'(l) = \sigma(l)$, $\forall l \in \mathcal{L}'$.

Often it is convenient to denote by L an integer matrix having the lattice $\mathcal{L}$ as its right image $\mathcal{L} := \left\{ Lm : m \in \mathbb{Z}^{\mathcal{E}} \right\}$. Thus, the columns of L span the lattice, and we abuse notation speaking of the partial character (L, σ) in this case. To each partial character $(\mathcal{L}, \sigma)$ we associate a *lattice ideal*:

$$I_+(\sigma) := \left\langle x^{m^+} - \sigma(m) x^{m^-} : m \in \mathcal{L} \right\rangle \subseteq \Bbbk[\mathcal{E}]. \tag{3.15}$$

It can be seen that a lattice ideal is prime if and only if its partial character is saturated, more generally, all associated primes of a lattice ideal arise from saturations of its partial character. A nice characterization is, that a proper binomial ideal $I \subseteq \Bbbk[x_1, \ldots, x_n]$ is a lattice ideal if and only if $I = (I : (\prod_{i=1}^n x_i)^{\infty})$. This fact can be used to compute a minimal generating set of a lattice ideal when only the partial character is given [HM09].

Now, cellular binomial ideals are a generalization of both lattice ideals and primary ideals. Radical cellular binomial ideals $I \subseteq \Bbbk[x_1, \ldots, x_n]$ are of the form $I = M(\mathcal{E}) + I_+(\sigma)$ for some partial character (L, σ) on $\mathbb{Z}^{\mathcal{E}}$. Now, assuming that $\Bbbk$ is algebraically closed, the associated primes of $M(\mathcal{E}) + I_+(\sigma)$ are given by

$$P_\tau = M(\mathcal{E}) + I_+(\tau), \tag{3.16}$$

where τ runs through all saturations of σ. If $\Bbbk$ is not algebraically closed, it may contain only some, or even no saturations of (L, σ). In Section 3.2.4 we give an algorithm that computes the minimal primes of a binomial ideal by directly computing a cellular decomposition of the radical of I into radical cellular ideals.

If the monomials in a cellular ideal I are of higher order, then we only have that $I \cap \Bbbk[\mathcal{E}]$ is a lattice ideal. However, the associated primes might have partial characters supported on different lattices. The key theorem for computing associated primes of cellular binomial ideals is

THEOREM 3.2.6 ([ES96], Theorem 8.1). *Let $I \subseteq \Bbbk[x_1, \ldots, x_n]$ be a cellular binomial ideal on the cell variables $\mathcal{E}$. Let $P = M(\mathcal{E}) + I_+(\sigma)$ be an associated prime of I, then there exists a monomial x^m in the variables not in $\mathcal{E}$ and a partial character τ on $\mathbb{Z}^{\mathcal{E}}$ whose saturation is σ, such that*

$$(I : x^m) \cap \Bbbk[\mathcal{E}] = I_+(\tau). \tag{3.17}$$

Note that the associated primes of a cellular binomial ideal are cellular binomial ideals for the same cell variables. To compute them, one only has to consider the quotients of I modulo the standard monomials in the variables outside $\mathcal{E}$. We want to point out that it is crucial to have a fast algorithm producing a cellular decomposition of a given binomial ideal. This is a necessary computation for finding the associated primes and also the primary components.

We now review an algorithm from [OP00] that is implemented in *Binomials*. It is based on the following approximation scheme for ideals in any Noetherian ring:

LEMMA 3.2.7 ([ES96], Proposition 7.2). *Let I be an ideal in a Noetherian ring S and $g \in S$ such that $(I : g) = (I : g^\infty)$. Then*

1. *$I = (I : g) \cap (I + \langle g \rangle)$.*
2. *$\operatorname{Ass}(S/(I : g)) \cap \operatorname{Ass}(S/(I + \langle g \rangle)) = \emptyset$.*
3. *A minimal primary decomposition of I consists of the primary components of $(I : g)$ and those primary components of $I + \langle g \rangle$ that correspond to associated primes of I.*

Given any noncellular binomial ideal I, we can find a variable x_i that is a zerodivisor but not nilpotent modulo I. A power $s > 0$ of that variable satisfies the conditions on g in Lemma 3.2.7 and we can write

$$I = (I : x_i^s) \cap (I + \langle x_i^s \rangle), \tag{3.18}$$

where the ideals on the right hand side are both binomial and properly containing I. This can be turned into a very simple algorithm for cellular decomposition, which was formulated in [OP00]. The authors also provided an implementation in Macaulay 2, parts of which are still used in the *Binomials* package.

ALGORITHM 3.2.8 (Cellular Decomposition). *Input: I, a binomial ideal.*
Output: A cellular decomposition of I.

1. *If I is cellular, return I.*
2. *Choose a variable that is a zerodivisor but not nilpotent modulo I.*
3. *Determine the power s such that $(I : x_i^s) = (I : x_i^\infty)$.*
4. *Iterate with $(I : x_i^s)$ and $I + \langle x_i^s \rangle$.*

Step 1 is carried out as follows. First determine the nilpotent variables by checking for which x_i one has $(I : x_i^\infty) = \Bbbk[x_1, \ldots, x_n]$. Denoting the remaining variables' indices as $\mathcal{E}$, I is cellular iff $\left(I : (\prod_{i \in \mathcal{E}} x_i)^\infty\right) = I$. Termination of Algorithm 3.2.8 is ensured since $\Bbbk[x_1, \ldots, x_n]$ is Noetherian and the two ideals $(I : x_i^s)$ and $I + \langle x_i^s \rangle$ properly contain I. Correctness follows from Lemma 3.2.7. Note that cellular components of pure difference binomial ideals are pure difference binomial ideals.

EXAMPLE 3.2.9 (Cellular Decomposition). We study an ideal from [ES96]. Let $S = \mathbb{Q}[x_1, \ldots, x_5]$ and $I = \langle x_1 x_4^2 - x_2 x_5^2, x_1^3 x_3^3 - x_4^2 x_2^4, x_2 x_4^8 - x_3^3 x_5^6 \rangle$.

```
i1 : S = QQ[x1,x2,x3,x4,x5];
i2 : I = ideal(x1*x4^2-x2*x5^2, x1^3*x3^3-x4^2*x2^4, x2*x4^8-x3^3*x5^6);
i3 : toString BCD I
o3 = {ideal(x1*x4^2-x2*x5^2,
            x1^3*x3^3-x2^4*x4^2,
            x2^3*x4^4-x1^2*x3^3*x5^2,
            x2^2*x4^6-x1*x3^3*x5^4,
            x2*x4^8-x3^3*x5^6),
      ideal(x1^2,x1*x4^2-x2*x5^2,
            x2^5,x5^6,x2^4*x4^2,x4^8)}
i4 : ap = binomialAssociatedPrimes I; toString ap
o4 = {ideal(x1*x4^2-x2*x5^2,
            x1^3*x3^3-x2^4*x4^2,
            x2^3*x4^4-x1^2*x3^3*x5^2,
            x2^2*x4^6-x1*x3^3*x5^4,
            x2*x4^8-x3^3*x5^6),
      ideal(x2,x5,x4,x1)}
i5 : intersect (ap#0,ap#1) == I
o5 = false
i6 : binomialRadical I == intersect (ap#0,ap#1)
o6 = true
i7 : isCellular (ap#0, returnCellVars=>true)
o7 = {x1, x2, x3, x4, x5}
i8 : isCellular (ap#1, returnCellVars=>true)
o8 = {x3}
```

In this listing we have suppressed some output. First we compute a cellular decomposition with `BCD`. It has two components. The first ideal is the toric ideal $(I : (\prod_{i=1}^{n} x_i)^\infty)$, which is prime. It is a general feature of the implementation of *Binomials* that, when the input has no monomial generators, the first ideal of the output of cellular and primary decompositions, as well as minimal and associated primes, is always the toric ideal. We also compute the associated primes. The second one is embedded, so the toric ideal is the only minimal prime. We confirm that I is not radical because of the monomial powers in the second component. Note also that the binomial generator in the second cellular component reduces to zero as soon as one takes the radicals of the monomials. Finally we confirm that the associated primes are cellular and show the set of variables with respect to which they are cellular, using `isCellular` with the option `returnCellVars`. The cell variables could have been computed directly together with the cellular decomposition by running the long version `binomialCellularDecomposition`, again with the option `returnCellVars` set to `true`.

Theorem 3.2.6 shows that saturation of partial characters is a crucial ingredient for computing the associated primes of a binomial ideal. We therefore study the properties of saturations of partial characters in the special case of pure difference binomial ideals. In the current implementation of *Binomials*, any operation that needs extension of the coefficient field of the polynomial ring is only implemented for pure difference binomial ideals. It will be shown that in this case cyclotomic field extensions suffice.

3.2.3. Solving Pure Difference Binomial Ideals. In this section we give a fast algorithm for solving pure difference binomial ideals of dimension zero. It is not surprising that such a procedure utilizes only the exponents of the generators. We denote by ξ_l the primitive l-th root of unity $\exp\left\{\frac{2\pi \mathrm{i}}{l}\right\} \in \mathbb{C}$. The field extension of $\mathbb{Q}$ that is obtained by adjoining such a root of unity is called a *cyclotomic field* and denoted by $\mathbb{Q}(\xi_l)$. It can be obtained constructively by taking the quotient of a univariate polynomial ring modulo the principal ideal generated by the minimal polynomial of ξ_l, the *cyclotomic polynomial* [Hun74, Chapter V].

PROPOSITION 3.2.10. *Given a zero-dimensional pure difference binomial ideal I, there exists a primitive root of unity ξ_l such that all complex solutions of I are contained in the cyclotomic field $\mathbb{Q}(\xi_l)$.*

The proof is given after the following Lemma, which is also of interest for the implementation.

LEMMA 3.2.11. *The complex solutions of the univariate equation*

$$x^n = \xi_m^k, \tag{3.19}$$

are given by the following roots of unity

$$x_0 = \xi_{mn}^k, \quad x_1 = \xi_{mn}^{m+k}, \quad \dots \quad x_{n-1} = \xi_{mn}^{(n-1)m+k}. \tag{3.20}$$

PROOF. The $x_0, \dots, x_{n-1}$ are n distinct roots of (3.19), which is of degree n. ☐

PROOF OF PROPOSITION 3.2.10. The standard method of reducing a multivariate problem to a univariate problem applies. The general framework is described for instance in Chapter 3 of [CLO96]. Choose an elimination term order, such as lexicographic order, and compute a Gröbner basis of I. This Gröbner basis consists of pure difference binomials since all S-polynomials are pure difference binomials. Furthermore, at least one of the binomials of this Gröbner basis is univariate as I is zero-dimensional and we have chosen an elimination order. The solutions of this univariate equation exist in a cyclotomic field by Lemma 3.2.11. We continue to extend the partial solution that we have found, substituting the variable for its value in the remaining elements of the Gröbner basis. We obtain a univariate equation in another variable. The final solution exists in the cyclotomic field containing all the roots of unity that are encountered in the course of the algorithm. ☐

Of course, the procedure that was just described is also valid for other fields $\Bbbk$. In the general case, field extensions have to be carried out by computing the minimal polynomial of the element to be adjoined and one has to do computations over the algebraic numbers. This however can become infeasible in practice since both the computations become lengthy and it becomes more and more tedious to produce output in a human-readable form.

We are now ready to formulate the algorithm for computing the variety of a zero-dimensional pure difference binomial ideal. The first thing that needs to be accounted for is the possibility of 0 as a solution, potentially with multiplicities. We take care of this by means of cellular decomposition. Each cellular binomial ideal I can be written as $I = \left((I + M(\mathcal{E})^d) : \left(\prod_{i\in\mathcal{E}} x_i\right)^\infty\right)$, and $I \cap \Bbbk[\mathcal{E}]$ is a lattice ideal. The solutions of I take the value zero at the variables outside $\mathcal{E}$ and each solution has a multiplicity of $\prod_{i\notin\mathcal{E}} d_i$.

ALGORITHM 3.2.12 (Solving pure difference binomial ideals).
Input: A zero-dimensional pure difference binomial ideal I.
Outputs: The root of unity that needs to be adjoined to $\mathbb{Q}$ and the list of the solutions of I.

(1) *Compute a cellular decomposition of I.*
(2) *For each cellular component:*

(a) *Set the noncell variables to zero, and determine the product* $D = \prod_{i \notin \mathcal{E}} d_i$ *of the powers of the noncell variables.*
(b) *Compute a lexicographic Gröbner basis and solve the lattice ideal of the cellular component, adjoining roots of unity where necessary.*
(c) *Save each solution* D *times.*

(3) *Compute the least common multiple* m *of the powers of the adjoined roots of unity and construct the cyclotomic field* $\mathbb{Q}(\xi_m)$*.*
(4) *Output the list of collected solutions as elements of* $\mathbb{Q}(\xi_m)$*.*

This algorithm is the main ingredient for saturating partial characters, which we treat after an example.

Example 3.2.13 (Solving Pure Difference Binomial Ideals). We solve a simple pure difference binomial ideal to introduce the syntax.

```
i1 : S = QQ[x,y,z];
i2 : I = ideal (x^2-y,y^3-z,x*y-z);

i3 : binomialSolve I
BinomialSolve created a cyclotomic field of order 3.

o3 = {{1, 1, 1}, {- ww_3 - 1, ww_3, 1}, {ww_3, - ww_3 - 1, 1},
     {0, 0, 0}, {0, 0, 0}, {0, 0, 0}}

i4 : degree I
o4 = 6
```

In the implementation, generic names consisting of `ww`, and the order, are assigned to roots of unity. Note that the square of the third root of unity `ww_3` is represented as `-ww_3-1` by means of its minimal polynomial over $\mathbb{Q}$. A cellular decomposition reveals that this ideal has two components, one of which is of degree 3 with associated prime $\langle x, y, z\rangle$. The function `binomialSolve` outputs the solutions with the correct multiplicities.

Saturations of partial characters exist only over algebraically closed fields. This is evident for instance from the partial character $((2), 2 \mapsto -1)$, consisting of the rank 1 lattice spanned by the integer 2, and the character that maps 2 to $-1 \in \mathbb{C}$. The saturations are pairs $(\mathbb{Z}, \tau)$, that satisfy $\tau(2) = \tau(1)^2 = -1$. This example is merely a combinatorial version of factorizing the polynomial $x^2 + 1$, which is the same as performing the primary decomposition of its principal ideal. The following algorithm to saturate a partial character is the general version of the example's principle.

Algorithm 3.2.14 (Saturation of a partial character).
Input: A partial character (L, σ)*, where* L *is a matrix whose columns are minimal generators of a lattice in* $\mathbb{Z}^d$*.*
Output: All distinct saturations $(\mathrm{Sat}(L), \tau_i)$*,* $i = 1, \dots, n$*.*

(1) *Compute the saturation* $L' := \mathrm{Sat}(L)$*, for instance by computing the minimal syzygies of the syzygies among the generators of* L*.*
(2) *Express the generators of* L *in terms of the generators of* L'*, by solving the matrix system*

$$L = L'K, \tag{3.21}$$

for the square matrix $K = (k_{ij})_{i,j=1,\dots,r}$*, where* $r := \mathrm{rk}(L) = \mathrm{rk}(L')$ *denotes the rank of the lattices.*

(3) Write l_j, l'_j, and k_j for the columns of L, L', and K, respectively. Introduce new variables $\tau_i := \tau(l'_i), i = 1, \ldots, r$, for the values that τ takes on the columns of L'. Using again monomial notation $\tau^m := \prod_{i=1}^r \tau_i^{m_i}$, compute the following zero-dimensional lattice ideal in $\mathbb{Q}[\tau_1, \ldots, \tau_r]$

$$J := \left(\left\langle \tau^{k_j^+} - \sigma(l_j)\tau^{k_j^-} : j = 1, \ldots, r \right\rangle : \Big(\prod_{i=1}^{r} \tau_i \Big)^\infty \right), \tag{3.22}$$

for the given values $\sigma(l_j)$.

(4) Solve J (over a suitable extension of $\mathbb{Q}$) and output L' together with the list of solutions of J.

Proof of correctness. Computing the saturation of a lattice should be viewed as an integer valued analogue of taking the orthogonal complement twice. The coefficient matrix K that solves the system (3.21) exists and is unique over $\mathbb{Z}$ as L is a sublattice of L' and we assumed that the columns of L' are a minimal set of generators of the corresponding lattice. The ranks of L and L' coincide by definition. The ideal J is constructed as follows: For each generator l of L we get a relation $l = L' \cdot k$, to which we apply the homomorphism τ, remembering that τ and σ are required to coincide on the generators of L. The entries of K are integers, thus we get the Laurent binomial ideal

$$\left\langle \sigma(l_j) - \prod_{i=1}^{r} \tau_i^{k_{ij}} : j = 1, \ldots, r \right\rangle, \tag{3.23}$$

whose intersection with $\mathbb{Q}[\tau_1, \ldots, \tau_r]$ is exactly J. That J is zero-dimensional follows since the quotient L'/L is a finite group. For details see Corollary 2.2 in [ES96]. □

We note without proof that the number of distinct saturations equals the order of the finite group $\mathrm{Sat}(L)/L$. Finally, for computing primary decompositions of pure difference binomial ideals we only need to solve such ideals during the saturation.

Proposition 3.2.15. *The saturation of a partial character that occurs during primary decomposition of a pure difference binomial ideal involves only solving pure difference binomial ideals.*

Proof. Any cellular component of a pure difference binomial ideal is pure difference again. So we can assume that I is cellular. Now, each partial character consists of a lattice and the constant map $l \mapsto 1$. Therefore the ideal J in Algorithm 3.2.14 is a pure difference binomial ideal. □

3.2.4. Minimal Primes of Binomial Ideals. In this section we will describe a new algorithm for computing the minimal primes of a binomial ideal. It is based on a variant of cellular decomposition, given in Algorithm 3.2.8. As we have seen previously the associated primes and thereby the minimal primes of a binomial ideal come in groups belonging to the cellular components of I. Our approach is to directly compute a cellular decomposition of the radical of I.

Algorithm 3.2.16 (Minimal primes of a binomial ideal).
Input: A binomial ideal $I \subseteq \Bbbk[x_1, \ldots, x_n]$.
Output: The binomial minimal primes of I.

(1) Determine whether I is cellular.

(a) *If yes, compute the radical* $(I \cap \Bbbk[\mathcal{E}]) + M(\mathcal{E}) = M(\mathcal{E}) + I_+(\sigma)$ *and its partial character* (L, σ). *Compute the saturations* $(\tau_i)_{i=1}^l$ *of* σ *and save the ideals*

$$P(\tau_i) = M(\mathcal{E}) + I_+(\tau_i). \tag{3.24}$$

(b) *If not, determine a variable* x_i *that is a zerodivisor, but not nilpotent modulo* I, *and iterate with the ideals* $I + \langle x_i \rangle$ *and* $(I : x_i^\infty)$.

(2) *From all primes collected, remove redundant ones to find a minimal prime decomposition of* $\operatorname{Rad}(I)$.

Proof of termination and correctness. Termination of this algorithm follows as the ambient ring is Noetherian and $I + \langle x_i \rangle$ and $(I : x_i^\infty)$ strictly contain I. The radical of I is the intersection of the ideals $I_{\mathcal{E}}$ in (4.2) of [ES96]. We encounter a decomposition of $\operatorname{Rad}(I)$ into such ideals in the course of the algorithm, as the iteration is ultimately producing cellular components of the radical of I. Thus, like in their Algorithm 9.2, correctness has been proved in Section 4 of [ES96]. For cellular ideals the minimal primes have the form (3.24), and the collection of all minimal primes of all cellular ideals contains the minimal primes of the original ideal by their Proposition 7.2. □

Remark 3.2.17. This algorithm differs from the cellular decomposition algorithm only in the recursion step, where we continue with $I + \langle x_i \rangle$ instead of $I + \langle x_i^s \rangle$. In this way we do not achieve a decomposition of I, but only of the radical of I. Fortunately, this algorithm can be significantly faster than cellular decomposition since adding variables, instead of higher powers of variables, allows the Gröbner basis engine to do more simplifications during the computation.

Example 3.2.18 (Binomial Minimal Primes). We continue where we left off in Example 3.2.9.

```
i16 : toString binomialMinimalPrimes I
o16 = {ideal(x1*x4^2-x2*x5^2,
             x1^3*x3^3-x2^4*x4^2,
             x2^3*x4^4-x1^2*x3^3*x5^2,
             x2^2*x4^6-x1*x3^3*x5^4,
             x2*x4^8-x3^3*x5^6)}
```

The result consists only of the toric ideal, confirming that the monomial prime is embedded. Although not visible from the output, the second associated prime was not computed on the way to this result, in particular the minimal primes are not extracted from a list of associated primes.

3.2.5. Primary Decomposition. The original primary decomposition algorithm in [ES96] was refined in [OP00]. The computation starts with a cellular decomposition, a first approximation of primary decomposition. In Eisenbud and Sturmfels' original paper, and also in [Alt00], there are cases identified in which a cellular decomposition is primary. If this is not the case, for each cellular component the associated primes need to be determined. Then finding the primary component can be achieved as follows. From an associated prime P of a cellular binomial ideal I, extract the "binomial part" $P^{(b)} = P \cap \Bbbk[\mathcal{E}]$. Then $I + P^{(b)}$ has P as its unique minimal prime. Computing the primary component over P is carried out by means of a localization operation called Hull, removing the embedded primary components of $I + P^{(b)}$. The refinement of [OP00] is to show that $I + P^{(b)}$ suffices in this procedure, while Eisenbud and Sturmfels originally suggested to add a sufficiently high monomial power. A combinatorial description of the resulting primary components is given in [DMM10], however, it seems difficult to use these results for computation.

A few remarks on primary decompositions in [ES96] and [OP00] are necessary. Corollary 6.5 of [ES96] shows that Hull(I) is a binomial ideal if I is a cellular binomial ideal. This Corollary is used in the proof of Theorem 7.1' to deduce that Hull(R_i) is binomial, where R_i is the sum of a monomial ideal and $I + P^{(b)}$ from above. However, it is not checked whether R_i is in fact cellular, as required by the Corollary. Example 3.2.19 shows a noncellular R_i that arises in the decomposition of the ideal of adjacent (2×2)-minors of a generic (5×5)-matrix. The computations necessary to check the example can be carried out easily with *Binomials*.

EXAMPLE 3.2.19. In the ring $\mathbb{Q}[a, b, \dots, o]$ consider the ideal

$$\begin{aligned} I = \big(&ln - ko, lm - jo, km - jn, l^2, kl, jl, k^2, jk, ik - hl, \\ &fk - cl, j^2, ij - gl, hj - gk, fj - al, cj - ak, fh - ci, \\ &fg - ai, cg - ah, f^2, cf, af, ce - bf, ae - df, c^2, ac, ab - cd, a^2\big). \end{aligned}$$

This ideal is cellular with respect to $\mathcal{E} = \{b, d, e, g, h, i, m, n, o\}$ and has four associated primes, which are pure difference. The binomial part of the unique minimal associated prime is

$$P^{(b)} = (in - ho, im - go, hm - gn).$$

Then $I + P^{(b)}$ has two cellular components whose sets of cell variables are $\mathcal{E}$ and $\{b, d, e, m, n, o\}$, respectively.

Using Theorem 7.1', in Algorithm 9.7 of [ES96] it is asked to compute Hull(R_i), using Algorithm 9.6. This however, requires a cellular ideal as its input. The algorithm can be corrected easily since the operation Hull is called only for ideals whose radical is prime. The associated primes of such an ideal have the radical as their unique minimal element, and as Hull removes embedded primary components, instead of Hull(R_i) we can compute Hull(Q_i) of any other ideal Q_i that has the same minimal prime. In particular we can choose $Q_i = \big(R_i : (\prod_{i\in\mathcal{E}} x_i)^\infty\big)$, the "cellularization" of R_i. Summarizing, in Algorithm 9.7, Step 3.3 should be replaced by

3.3' Compute Hull $\big(R_i : \big(\prod_{i\in\mathcal{E}} x_i\big)^\infty\big)$ using Algorithm 9.6.

Unfortunately, also in Theorem 3.2 of [OP00], Corollary 6.5 of [ES96] is used to deduce that Hull($I + (P \cap \Bbbk[\mathcal{E}])$) is binomial and primary. Again, this is wrong as $I + (P \cap \Bbbk[\mathcal{E}])$ is not cellular. The result can be saved by first cellularizing as explained above. The implementation in *Binomials* incorporates these modifications and is demonstrated next.

EXAMPLE 3.2.20 (Binomial Primary Decomposition). We compute the primary decomposition of $I = \langle x^2 - y, y^2 - z, z^2 - x\rangle \in \mathbb{Q}[x, y, z]$.

```
i1 : S = QQ[x,y,z]
i2 : I = ideal(x^2-y,y^2-z^2,z^2-x)
i3 : dim I
o3 = 0
i4 : degree I
o4 = 8

i5 : bpd = BPD I
Running cellular decomposition:
cellular components found: 1
```

```
cellular components found: 2
Decomposing cellular component: 1 of 2
BinomialSolve created a cyclotomic field of order 6.
done
Decomposing cellular component: 2 of 2
done
Removing redundant components (fast)

o6 = {ideal(z+ww_6-1,y-ww_6+1,x+ww_6),
      ideal(z+ww_6,y+ww_6,x-ww_6+1), ideal(z+1,y-1,x-1),
      ideal(z-1,y-1,x-1), ideal(z-ww_6,y+ww_6,x-ww_6+1),
      ideal(z-ww_6+1,y-ww_6+1,x+ww_6), ideal(y,x,z^2)}

i7 : intersect bpd == sub (I, ring bpd#0)
o7 = true
```

The function `BPD` is only a convenient shorthand for `binomialPrimaryDecomposition`, which can also be used in the long form and offers some options. The primary decomposition of I into binomial ideals exists in $\mathbb{Q}(\xi_6)[x, y, z]$, so `BPD` created this cyclotomic field, calling the primitive sixth root of unity `ww_6`. Observe that the ideal has a double zero at the origin. In `i7` we intersect the result to confirm that the decomposition is correct. The result of the intersection is defined over the extended polynomial ring $\mathbb{Q}(\xi_6)[x, y, z]$, and can be compared to I only after mapping it to that ring.

This concludes our overview of the functionality of *Binomials* and we move on to the discussion of some large primary decompositions.

3.3. A nonradical Commuting Birth and Death Ideal

In this section we study the commutative algebra of discrete time commuting birth and death ideals. One-dimensional birth and death processes are among the simplest Markov chains that are considered in modeling random processes [LR99]. In the discrete time case, many of their properties can be derived from the explicit spectral theory of transition matrices. The paper [ESU10] gives motivation to consider generalized processes that correspond to Markov chains on multi-dimensional lattices, and as most of the one-dimensional theory does not apply there, the authors strive to identify subclasses with nice properties. The work suggests commuting birth and death processes which are defined by transition matrices having the property that transitions in the different dimensions commute. After reformulation, these conditions can be seen to result in binomial conditions on the entries of the transition matrices, that is, a binomial ideal. The toric component of this binomial ideal nicely relates to an underlying matroid as discussed in the paper. Determining primary decompositions of commuting birth and death ideals poses interesting challenges in combinatorial commutative algebra.

Computational results given in this section tend to be very large. We have therefore stored them on a web page, which also contains additional scripts to reproduce the results:

(3.25) `http://personal-homepages.mis.mpg.de/kahle/cbd/`

We now define the binomial ideals under consideration. The ambient polynomial ring has indeterminates corresponding to the edges of a regular grid. For fixed integers

$n_1, \ldots, n_m$, let

$$E := \prod_{i=1}^{m} \{0, \ldots, n_i - 1\}, \tag{3.26}$$

be the usual m-dimensional bounded regular grid with edges between vertices that differ by ± 1 in exactly one coordinate. Here it is sufficient to consider only the cases $m = 2, 3$. For each edge in the grid we define two indeterminates, one for each direction. In the two-dimensional case the authors used the notation $\Bbbk[R, L, D, U]$ to denote a polynomial ring in the indeterminates

$$\begin{aligned} &\{R_{ij} : 0 \le i < n_1, 0 \le j \le n_2\} \cup \{L_{ij} : 0 < i \le n_1, 0 \le j \le n_2\} \cup \\ &\{D_{ij} : 0 \le i \le n_1, 0 < j \le n_2\} \cup \{U_{ij} : 0 \le i \le n_1, 0 \le j < n_2\}, \end{aligned} \tag{3.27}$$

where R_{ij} is supposed to represent a *right* move starting at position ij and so on. In the case $m = 3$ one can, in a natural way, extend the set of indeterminates by introducing letters F and B and three indices for each indeterminate. The set of commuting birth and death processes is defined by the binomial equations (3.1) of [ESU10]. These equations arise in quadruples, coming from squares in the graph E, by which we mean induced subgraphs G of E that are isomorphic to the usual square. Denoting its vertices by $\{(u, v), (u + e_i, v), (u, v + e_j), (u + e_i, v + e_j)\}$, the corresponding ideal encodes that the two paths joining opposite vertices are equivalent:

(3.28)

$$\begin{aligned} I^G := \Big\langle & U_{(u,v)} R_{(u,v+e_j)} - R_{(u,v)} U_{(u+e_i,v)}, \quad D_{(u,v+e_j)} R_{(u,v)} - R_{(u,v+e_j)} D_{(u+e_i,v+e_j)}, \\ & L_{(u+e_i,v+e_j)} D_{(u,v+e_j)} - D_{(u+e_i,v+e_j)} L_{(u+e_i,v)}, \quad L_{(u+e_i,v)} U_{(u,v)} - U_{(u+e_i,v)} L_{(u+e_i,v+e_j)} \Big\rangle. \end{aligned}$$

The commuting birth and death ideal is the sum of all I^G, where G runs through the induced squares of E.

$$I^E := \sum_{G \text{ square in } E} I^G. \tag{3.29}$$

In the case $m = 2, 3$ these ideals have been denoted $I^{(n_1,n_2)}$, and $I^{(n_1,n_2,n_3)}$ in [ESU10].

EXAMPLE 3.3.1. The graph E for $m = 2$ and $n_1 = n_2 = 1$ is just a square and $I^{(1,1)}$ is generated by the four binomials

$$\begin{aligned} I^{(1,1)} = \langle & U_{00}R_{01} - R_{00}U_{10}, \quad R_{01}D_{11} - D_{01}R_{00}, \\ & D_{11}L_{10} - L_{11}D_{01}, \quad L_{10}U_{00} - U_{10}L_{11}\rangle. \end{aligned} \tag{3.30}$$

If $m = 3$ and $n_1 = n_2 = n_3 = 1$, E is the 3-cube and the squares arise from facets. Thus, $I^{(1,1,1)}$ is generated by 24 pure difference binomials, 4 for each facet.

On the web page (3.25) one can download Python scripts that generate Macaulay 2 code for the rings and ideals in the cases $m = 2, 3$. The following shows an example how to use the script `Imn.py` on the command line to generate $I^{(2,2)}$:

```
> ./Imn.py 2 2
-- Macaulay 2 Code for the Commuting Birth and Death Ideal:
-- m = 2, n = 2
S = QQ[R00,U00,R01,D01,U01,R02,D02,R10,L10,U10,R11,L11,D11,U11,
       R12,L12,D12,L20,U20,L21,D21,U21,L22,D22];
I = ideal
(U00*R01-R00*U10,R01*D11-D01*R00,D11*L10-L11*D01,L10*U00-U10*L11,
 U01*R02-R01*U11,R02*D12-D02*R01,D12*L11-L12*D02,L11*U01-U11*L12,
```

```
U10*R11-R10*U20,R11*D21-D11*R10,D21*L20-L21*D11,L20*U10-U20*L21,
U11*R12-R11*U21,R12*D22-D12*R11,D22*L21-L22*D12,L21*U11-U21*L22);
```

In [ESU10] the authors discuss the primary decompositions of $I^{(2,2)}$, $I^{(1,1,1)}$, and smaller examples. They state that these computations could not be carried out with the standard implementations, but were derived in an interactive session. The current implementation of *Binomials* computed the 199 prime components of $I^{(2,2)}$ in 100 seconds and took 123 seconds to decompose $I^{(1,1,1)}$ on the author's 1,6 GHz laptop. As mentioned before, computing the minimal primes directly is even faster and can be completed in half of the time.

Based on their results, Evans, Sturmfels, and Uhler conjectured

CONJECTURE 3.3.2. For any grid E, the ideal I^E is radical, its prime decomposition consists of pure toric ideals and is independent of the coefficient field.

Here a *pure toric ideal* is an ideal generated by indeterminates and pure difference binomials. In [ESU10] the authors prove that every associated prime of $I^{(1,n)}$ is a pure toric ideal.

THEOREM 3.3.3. *The ideal $I^{(2,3)}$ is the intersection of 2638 primary binomial ideals whose properties are given in Table 1. Among these are 10 components that are not prime, and thus $I^{(2,3)}$ is not radical. The* 10 *associated primes of these components are all embedded and of codimension 20. The radical* $\mathrm{Rad}(I^{(2,3)})$ *is the intersection of* 2628 *minimal primes and given by the following ideal:*

$$
\begin{aligned}
I^{(2,3)} + \langle & D_{01}R_{03}R_{10}L_{12}U_{21}L_{22}D_{23} - U_{01}R_{03}L_{10}R_{13}D_{21}L_{23}D_{23},\\
& U_{00}R_{02}R_{12}L_{13}L_{20}D_{22}U_{22} - R_{00}D_{02}R_{13}L_{13}U_{20}U_{22}L_{23},\\
& R_{00}U_{01}R_{03}L_{10}R_{13}U_{20}L_{23}D_{23} - U_{01}R_{03}^2R_{13}L_{13}U_{20}L_{23}D_{23},\\
& R_{00}D_{02}L_{10}R_{13}L_{13}D_{21}U_{22}L_{23} - D_{02}R_{03}R_{13}L_{13}^2D_{21}U_{22}L_{23},\\
& U_{00}R_{02}R_{03}R_{12}L_{13}L_{20}D_{22}D_{23} - R_{00}D_{02}R_{03}R_{13}L_{13}U_{20}L_{23}D_{23},\\
& U_{00}R_{03}R_{10}L_{12}L_{20}U_{21}L_{22}D_{23} - U_{00}R_{03}L_{12}R_{13}U_{21}L_{22}L_{23}D_{23},\\
& R_{00}D_{03}L_{11}R_{13}U_{20}L_{21}D_{22}L_{23} - U_{00}R_{03}L_{12}R_{13}L_{20}L_{22}D_{22}D_{23},\\
& R_{01}U_{02}L_{10}R_{11}R_{13}D_{21}U_{21}L_{23} - D_{01}R_{02}R_{10}R_{12}L_{13}U_{21}U_{22}L_{23},\\
& D_{01}R_{02}R_{10}R_{12}L_{13}L_{20}D_{22}U_{22} - D_{01}R_{02}R_{12}R_{13}L_{13}D_{22}U_{22}L_{23},\\
& D_{01}R_{03}R_{10}L_{12}L_{13}U_{21}L_{22}U_{22} - U_{01}R_{03}L_{10}R_{13}L_{13}D_{21}U_{22}L_{23}\rangle.
\end{aligned}
\tag{3.31}
$$

One should note the two squares of variables in the third and fourth generator of $\mathrm{Rad}(I^{(2,3)})$. To produce these results one can use the functions `BPD` and `binomialMinimalPrimes`. The author's computer determined the minimal primes in approximately 4 hours. Taking the intersection of these primes took another hour on a 2,8 GHz AMD Opteron. Care has to be taken when computing intersections of many primes. In Macaulay 2 versions 1.2 and below, using the command `intersect` directly on a large list of primes will not terminate. If one does the intersection manually with a loop, intersecting only two ideals at a time, everything is fine. Computing the cellular and primary decomposition was more delicate. It took several days and used about 5 GB of RAM. In fact, the original computation of the cellular decomposition was done with a slightly different algorithm which only works if the toric component is isolated. We first computed the toric component T independently with the tool 4ti2 and then removed it by computing the saturation $(I^{(2,3)} : T^{\infty})$. The cellular decomposition of this ideal was easier to compute. Surprisingly this is not always the case. For some ideals I, with toric component T, the saturation $(I : T^{\infty})$ is just too complicated to be

codimension	16	17	17	18	18	19	19	20	20	21	21	22	22
# of components	1	14	2	107	91	356	612	527	550	212	120	38	8
gen. max degree	1	1	4	1	6	1	5	1	4	1	2	1	3
degree	1	1	64	1	4012	1	144	1	36	1	12	1	3
monomial	y	y	n	y	n	y	n	y	n	y	n	y	n

TABLE 1. Statistics on the primary components of $I^{(2,3)}$ sorted by codimension. Monomial components have been separated from binomial ones as indicated in the row "monomial". The row "gen. max degree" gives the maximal degree of a generator in this codimension while "degree" refers to the maximal degree among components. The toric component is generated in degree 6, of codimension 18 and degree 4012.

n	*1*	*2*	*3*	*4*	*5*	*6*
# of components	*3*	*11*	*40*	*139*	*466*	*1528*

TABLE 2. Prime decompositions of $I^{(1,n)}$

computed with Macaulay 2. In some cases, simply doing the cellular decomposition with Algorithm 3.2.8 is faster.

To complete this computational study, we have also investigated the ideals $I^{(1,n)}$ for $n \leq 6$. It was not possible to find a counterexample there.

THEOREM 3.3.4. *The ideals* $I^{(1,n)}$, $n = 1, \ldots, 6$ *are radical. The respective numbers of prime components are given in Table 2.*

Concluding this section we find that the conjecture turned out to be false in full generality. It might however hold for the ideals $I^{(1,n)}$, and the associated primes could still be pure toric ideals for all I^E.

Outlook and Conclusion

Discrete Exponential Families and Hierarchical Models

The support set problem for discrete exponential families is to characterize the face lattice of the convex support. We have exposed the combinatorial nature of this problem by tracing it back to an underlying realizable oriented matroid. In this way an exponential family is just another representation of this oriented matroid, making its relation to the convex support completely natural. We have shown that the implicit representation of a statistical model follows from the oriented matroid. This fact holds also in the general case of an arbitrary sufficient statistics, while a Markov basis description is only available in special cases. It is often underestimated in algebraic statistics, that to describe a toric variety a Markov basis is necessary. However, the circuits suffice to describe a statistical model, the nonnegative real part of the variety. The view from oriented matroids provides different perspectives of exponential families and will hopefully contribute to further understanding of structural zeros. Its main use might be the availability of a broad variety of examples constructed in combinatorial theory.

For hierarchical models, elementary circuits have been constructed. We have seen which of them originate in conditional independence statements and constructed hierarchical models whose Markov bases consist of elementary circuits. The derivation of bounds on supports of Markov moves immediately gives the neighborliness property of the marginal polytope. We have investigated the structure of binary marginal polytopes and exposed their connection to the theory of linear codes. This work could be continued by investigating the face lattices of marginal polytopes. For the neighborliness we identified equations that are necessarily fulfilled by distributions with small support. Coming from the other side, one could try to generate sets of vertices that do not form faces by looking at violations of equations that hold on the model. Even if the complete set of circuits is unknown, elementary circuits and their linear combinations are a source of such equations. Then the symmetry has to be incorporated. For instance in the binary k-interaction models one has the semidirect product $S_n \ltimes \mathbb{Z}_2^n$ acting on the marginal polytope, where S_n permutes the random variables while each copy of $\mathbb{Z}_2$ exchanges the roles of the states 0 and 1 in a given random variable. As some marginal polytopes are CUT-polytopes, it seems plausible that the methods described in [DL97] could be generalized to marginal polytopes.

Binomial Ideals

We have discussed decompositions of binomial ideals and shown an implementation of specialized algorithms for this purpose. Many operations on binomial ideals have been translated to operations on exponent vectors, or on the associated partial characters. By "making them combinatorial" significant speed-ups can be achieved. We have disproved a recent conjecture of [ESU10] by computing one of the largest primary decompositions ever. In future research it seems feasible to solve the case for the ideals $I^{(1,n)}$ by proving the conjecture there.

A natural continuation of the theoretical part of this work is to investigate decompositions that are finer than cellular decompositions, but not as fine as primary decompositions. For applications it is often useful to have a finest decomposition into pure difference binomial ideals, which is in particular independent of the coefficient field. Consider for instance the pure difference binomial ideal $I = \langle x^{17} - 1 \rangle \subseteq \mathbb{C}[x]$. Since $\mathbb{C}$ is algebraically closed, I admits a primary decomposition into binomial ideals, equivalent to factoring $x^{17} - 1$. In an application however, we might wish to avoid further decomposition. In a way, the remaining computation is trivial and would only clutter the output. The special combinatorial structure of (pure difference) binomial ideals indicates a way how to achieve coarser decompositions. The associated primes of a cellular binomial ideal can be grouped according to integer lattices supporting them (see Theorem 3.2.6). In this grouping all the 17 associated primes of I are supported on the same lattice. Informally, different associated primes supported on the same integer lattice indicate that the primary decomposition requires factorizing univariate polynomials, a step that we want to avoid. This connection also shows how to solve the problem. A *mesoprimary* ideal should be an ideal whose associated primes are supported on a single lattice. Then the mesoprimary components of an ideal are binomial by Theorem 6.4 of [ES96]. Results in this direction will eventually allow a clean separation of the combinatorial decomposition, which is related to decompositions of congruences on semigroup rings, and the field dependence arising from saturation of partial characters. This is the subject of the author's ongoing research.

Computation

In applications, and also for generating intuition, explicit computation is extremely useful. The field of computational algebra sees constant exchange with applications, commutative algebra, and also algebraic geometry. The implementation of specialized algorithms for monomial and binomial ideals, exploiting their combinatorial structure, lead to significant speed-ups and enables to tackle problems that are infeasible with general purpose implementations.

For the package *Binomials*, one next step is to complete the implementation of binomial primary decomposition over finite fields. Also for pure difference ideals in characteristic zero there is room for improvement. A very frequent computation in binomial primary decomposition is saturation with respect to monomials. Any implementation will immediately gain speed if the following problem was solved

PROBLEM 3.3.5. *Develop a specialized algorithm to compute, for any (cellular) binomial ideal I, the "partially saturated" ideal*

$$I : \left(\prod_{i \in \mathcal{E}} x_i \right)^{\infty} . \tag{3.32}$$

The software 4ti2 implements the *project-and-lift* algorithm, a fast algorithm for computing the saturation

$$I : \left(\prod_{i=1}^{n} x_i \right)^{\infty} . \tag{3.33}$$

It seems natural to extended the program to solve the above problem. The implementation of algorithms for binomial ideals presented here will greatly benefit from this feature and *Binomials* is prepared to incorporate it upon availability.

Bibliography

[4ti207] 4ti2 team, *4ti2—a software package for algebraic, geometric and combinatorial problems on linear spaces*, available at www.4ti2.de, 2007.

[AK06] N. Ay and A. Knauf, *Maximizing multi-information*, Kybernetika **42** (2006), 517–538.

[AKN92] S. Amari, K. Kurata, and H. Nagaoka, *Information geometry of Boltzmann machines*, IEEE Transactions on Neural Networks **3** (1992), no. 2, 260–271.

[Alt00] K. Altmann, *The chain property for the associated primes of $\mathcal{A}$-graded ideals*, Mathematical Research Letters **7** (2000), 565–575.

[Ama00] S. Amari, *Methods of information geometry*, Translations of mathematical Monographs, vol. 191, American Mathematical Society, 2000.

[Ama01] ———, *Information geometry on hierarchy of probability distributions*, IEEE Transactions on Information Theory **47** (2001), 1701–1711.

[AT03] S. Aoki and A. Takemura, *The list of indispensable moves of the unique minimal Markov basis for $3\times4\times k$ and $4\times4\times4$ contingency tables with fixed two-dimensional marginals*, METR 2003-38, University of Tokyo, Japan, 2003.

[Ay02] N. Ay, *An information-geometric approach to a theory of pragmatic structuring*, Annals of Probability **30** (2002), 416–436.

[BJT07] T. Bogarta, A.N. Jensen, and R.R. Thomas, *The circuit ideal of a vector configuration*, Journal of Algebra **309** (2007), no. 2, 518–542.

[BN78] O. Barndorff-Nielsen, *Information and exponential families in statistical theory*, Wiley, 1978.

[BV88] W. Bruns and U. Vetter, *Determinantal rings*, LNM, vol. 1327, Springer Verlag, Berlin, 1988.

[BVS$^+$93] A. Björner, M.L. Vergnas, B. Sturmfels, N. White, and G.M. Ziegler, *Oriented matroids*, Encyclopedia of Mathematics and its Applications, vol. 46, Cambridge University Press, Cambridge, 1993.

[Cai75] E. Caianiello, *Synthesis of boolean nets and time behaviour of a general mathematical neuron*, Biol. Cybernetics **18** (1975), 111–117.

[Cai86] ———, *Neuronic equations revisited and completely solved*, Brain Theory (G. Palm and A. Aertsen, eds.), Springer, Berlin, 1986.

[Chr97] R. Christensen, *Log-linear models and logistic regression*, second ed. ed., Springer Texts in Statistics, Springer-Verlag, New York, 1997.

[CLO96] D.A. Cox, J.B. Little, and D. O'Shea, *Ideals, varieties, and algorithms*, second ed., UTM, Springer, New York, 1996.

[CLS09] D. Cox, J. Little, and H. Schenck, *Toric varieties*, Manuscript of forthcoming book, 2009.

[CM03] I. Csiszár and F. Matúš, *Information projections revisited*, IEEE Transactions on Information Theory **49** (2003), 1474–1490.

[CM04] I. Csiszár and F. Matúš, *On information closures of exponential families: A counterexample*, IEEE Trans. Inform. Th. **50** (2004), no. 5, 922–924.

[CM05] ———, *Closures of exponential families*, The Annals of Probability **33** (2005), 582–600.

[CoC] CoCoATeam, CoCoA*: a system for doing Computations in Commutative Algebra*, Available at `http://cocoa.dima.unige.it`.

[CS04] I. Csiszár and P.C. Shields, *Information theory and statistics: A tutorial*, Foundations and Trends in Communications and Information Theory, now Publishers, 2004.

[DL97] M. Deza and M. Laurent, *Geometry of cuts and metrics*, Algorithms and Combinatorics, Springer, Berlin, 1997.

[DMM10] A. Dickenstein, L. Matusevich, and E. Miller, *Combinatorics of binomial primary decomposition*, Mathematische Zeitschrift **264** (2010), no. 4, 745–763.

[DS98] P. Diaconis and B. Sturmfels, *Algebraic algorithms for sampling from conditional distributions*, Annals of Statistics **26** (1998), 363–397.

[DS03] M. Develin and S. Sullivant, *Markov bases of binary graph models*, Annals of Combinatorics **7** (2003), 441–466.

[DSL04] G. Dósa, I. Szalkai, and C. Laflamme, *The maximum and minimum number of circuits and bases of matroids*, Pure Mathematics and Applications **15** (2004), no. 4, 383–392.

[DSS09] M. Drton, B. Sturmfels, and S. Sullivant, *Lectures on algebraic statistics*, Oberwolfach Seminars, vol. 39, Springer, Berlin, 2009, A Birkhäuser book.

[EGSS01] D. Eisenbud, D.R. Grayson, M.E. Stillman, and B. Sturmfels, *Computations in algebraic geometry with Macaulay 2*, Algorithms and Computations in Mathematics, no. 8, Springer-Verlag, New York, 2001.

[EH01] D. Eisenbud and J. Harris, *The geometry of schemes*, Springer-Verlag, New York, 2001.

[Eis95] D. Eisenbud, *Commutative algebra: with a view toward algebraic geometry*, GTM, vol. 150, Springer Verlag, New York, 1995.

[Eng08] A. Engström, *Cut ideals of K4-minor free graphs are generated by quadrics*, 2008, arXiv:0805.1762.

[ES96] D. Eisenbud and B. Sturmfels, *Binomial ideals*, Duke Mathematical Journal **84** (1996), no. 1, 1–45.

[ESU10] S. Evans, B. Sturmfels, and C. Uhler, *Commuting birth-death processes*, Annals of Applied Probability **20** (2010), no. 1, 238–266.

[Fin09] A. Fink, *The binomial ideal of the intersection axiom for conditional probabilities*, preprint arXiv:0902.1495 (2009).

[Ful93] W. Fulton, *Introduction to toric varieties*, Annals of Mathematical Studies, Princeton University Press, 1993.

[Geo88] H.O. Georgii, *Gibbs measures and phase transitions*, de Gruyter Studies in Mathematics, no. 9, de Gruyter, Berlin, 1988.

[Gil84] R. Gilmer, *Commutative semigroup rings*, University of Chicago Press, Chicago, 1984.

[GMS06] D. Geiger, C. Meek, and B. Sturmfels, *On the toric algebra of graphical models*, The Annals of Statistics **34** (2006), no. 5, 1463–1492.

[GS] D.R. Grayson and M.E. Stillman, *Macaulay 2, a software system for research in algebraic geometry*, Available at `http://www.math.uiuc.edu/Macaulay2/`.

[HHH+10] J. Herzog, T. Hibi, F. Hreinsdóttir, T. Kahle, and J. Rauh, *Binomial edge ideals and conditional independence statements*, Advances in Applied Mathematics (2010), in press, doi:10.1016/j.aam.2010.01.003.

[HM09] R. Hemmecke and P. Malkin, *Computing generating sets of lattice ideals*, Journal of Symbolic Computation **44** (2009), no. 10, 1493–1476.

[HMS$^+$08] R. Hemmecke, J. Morton, A. Shiu, B. Sturmfels, and O. Wienand, *Three counter-examples on semi-graphoids*, Combinatorics, Probability and Computation **17** (2008), 239–257.

[HS02] S. Hoşten and S. Sullivant, *Gröbner bases and polyhedral geometry of reducible and cyclic models*, Journal of Combinatorial Theory: Series A **100** (2002), no. 2, 277–301.

[HS04] ———, *Ideals of adjacent minors*, Journal of Algebra **277** (2004), no. 2, 615–642.

[Hun74] T.W. Hungerford, *Algebra*, GTM, vol. 73, Springer-Verlag, New York, 1974.

[KA06] T. Kahle and N. Ay, *Support sets of distributions with given interaction structure*, Proceedings of the WUPES 2006, 2006.

[Kah06] T. Kahle, *Information geometry and statistical learning theory*, Master's thesis, University of Leipzig, 2006.

[Kah10a] ———, *Decompositions of binomial ideals*, Ann. Inst. Stat. Math. **62** (2010), no. 4, 727–745.

[Kah10b] ———, *Neighborliness of marginal polytopes*, Contributions to Algebra and Geometry **51** (2010), no. 1, 45–56.

[KL51] S. Kullback and R.A. Leibler, *On information and sufficiency*, Annals of Mathematical Statistics **22** (1951), no. 1, 79–86.

[KWA09] T. Kahle, W. Wenzel, and N. Ay, *Hierarchical models, marginal polytopes, and linear codes*, Kybernetika **45** (2009), no. 2, 189–208.

[Lau96] S.L. Lauritzen, *Graphical models*, Oxford Statistical Science Series, Oxford University Press, 1996.

[LO06] J.A.D. Loera and S. Onn, *Markov bases of three-way tables are arbitrarily complicated*, Journal of Symbolic Computation **41** (2006), 173–181.

[LR99] G. Latouche and V. Ramaswami, *Introduction to matrix analytic methods in stochastic modeling*, Statistics and Applied Probability, SIAM, Philadelphia, 1999.

[MA03] F. Matúš and N. Ay, *On maximization of the information divergence from an exponential family*, Proceedings of WUPES'03 (J. Vejnarova, ed.), 2003, pp. 199–204.

[Mal06] P.N. Malkin, *Truncated Markov bases and Gröbner bases for integer programming*, 2006.

[Mat07] F. Matúš, *Optimality conditions for maximizers of the information divergence from an exponential family*, Kybernetika **43** (2007), no. 5, 731–746.

[McM70] P. McMullen, *The maximum number of faces of a convex polytope*, Mathematika (1970), 179–184.

[MR93] M.K. Murray and J.W. Rice, *Differential geometry and statistics*, Chapman & Hall, London, 1993.

[MS05] E. Miller and B. Sturmfels, *Combinatorial commutative algebra*, GTM, vol. 227, Springer, Berlin, 2005.

[OP00] I. Ojeda and R. Piedra, *Cellular binomial ideals. primary decomposition of binomial ideals*, Journal of Symbolic Computation **30** (2000), 383–400.

[O'S09] E. O'Shea, *On the occurrence of large gaps in small contingency tables*, arXiv:0912.0702 (2009).

[Pon66] L.S. Pontryagin, *Topological groups*, second edition ed., Gordon and Breach, New York, London, Paris, 1966.

[PRW01] G. Pistone, E. Riccomagno, and H.P. Wynn, *Algebraic statistics*, Monographs on Statistics and Applied Probability, Chapman & Hall CRC, 2001.

[PS05] L. Pachter and B. Sturmfels (eds.), *Algebraic statistics for computational biology*, Cambridge University Press, 2005.

[Rau09] J. Rauh, *Finding the maximizers of the information divergence from an exponential family*, arxiv:0912.4660 (2009).

[RK] J. Rauh and T. Kahle, *The Markov bases database*, http://mbdb.mis.mpg.de.

[RKA09] J. Rauh, T. Kahle, and N. Ay, *Support sets in exponential families and oriented matroid theory*, Proceedings of the WUPES'09, 2009.

[SJ08] D. Sontag and T. Jaakkola, *New outer bounds on the marginal polytope*, Advances in Neural Information Processing Systems 20 (Cambridge, MA) (J. Platt, D. Koller, Y. Singer, and S. Roweis, eds.), MIT Press, 2008, pp. 1393–1400.

[SK08] L. Steiner and T. Kahle, *cipi—computing information projections iterativly*, Available at http://personal-homepages.mis.mpg.de/kahle/cipi, 2008.

[Sta96] R.P. Stanley, *Combinatorics and commutative algebra*, 2nd ed., Birkhäuser, 1996.

[Stu88] B. Sturmfels, *Some applications of affine gale diagrams to polytopes with few vertices*, SIAM J. Discrete Mathematics **1** (1988), 121–133.

[Stu96] ——, *Gröbner bases and convex polytopes*, University Lecture Series, vol. 8, American Mathematical Society, Providence, RI, 1996.

[Stu02] ——, *Solving systems of polynomial equations*, CBMS, vol. 97, American Mathematical Society, 2002.

[TA02] A. Takemura and S. Aoki, *Some characterizations of minimal Markov basis for sampling from discrete conditional distributions*, Annals of the Institute of Statistical Mathematics, 2002, pp. 1–17.

[Tea] GraphViz. Team, *The dot language*, `http://www.graphviz.org/doc/info/lang.html`.

[van99] J. van Lint, *Introduction to coding theory*, third edition ed., GTM, Springer, Berlin, 1999.

[Wen06] W. Wenzel, *Regular simplices inscribed into the cube and exhibiting a group structure*, Journal of Combinatorial Mathematics and Combinatorial Computing **59** (2006), 213–220.

[Win03] G. Winkler, *Image analysis, random fields and Markov Chain Monte Carlo methods*, 2nd ed., Springer, 2003.

[WJ03] M.J. Wainwright and M.I. Jordan, *Variational inference in graphical models: The view from the marginal polytope.*, Allerton Conference on Communication, Control, and Computing, 2003.

[Zie94] G.M. Ziegler, *Lectures on polytopes*, GTM, vol. 152, Springer Verlag, 1994.

Index

Daten zum Autor

Name:	Thomas Kahle
Geburtsdatum:	17. Dezember 1981 in Magdeburg
10/2001-09/2003	Studium der Physik Universität Magdeburg
10/2003-09/2006	Studium der Physik Universität Leipzig
10/2006-05/2010	Doktorand am Max-Planck-Institut für Mathematik in den Naturwissenschaften, Leipzig

Bibiliographische Daten

On Boundaries of Statistical Models
(Randeigenschaften Statistischer Modelle)
Kahle, Thomas
Universität Leipzig, Dissertation, 2009

83 Seiten, 1 Abbildung, 82 Referenzen

Selbstständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichen oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialien oder erbrachten Dienstleistungen als solche gekennzeichnet.

Leipzig, den 15. Dezember 2009

. . . .

(Thomas Kahle)