Finding the Maximizers of the Information Divergence from an Exponential Family

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 Dipl.-Math. Dipl.-Phys. Johannes Rauh,
geboren am 30. September 1981 in Würzburg.

Die Annahme der Dissertation wurde empfohlen von:

1. Professor Dr. Jürgen Jost (MPI MIS Leipzig)
2. Professor Dr. Andreas Knauf (Universität Erlangen-Nürnberg)

Die Verleihung des akademischen Grades erfolgt mit Bestehen der
Verteidigung am 1.9.2011
mit dem Gesamtprädikat *summa cum laude*.

Contents

1. Introduction

Motivation and previous results. The subject of the present thesis is the maximization of the information divergence $D_{\mathcal{E}}(P) := D(P\|\mathcal{E})$ from an exponential family $\mathcal{E}$ over a finite set. This problem was first formulated by Nihat Ay. The original motivation was the quest for global variational principles that explain local learning rules in neural networks, in particular Hebb's rule. One such principle is the infomax principle, suggested by Linsker [47] in 1988. Later, in 2002, Ay suggested a variation, the IMI principle [6]. Both principles stipulate that a learning neural network tries to maximize the mutual information or multiinformation between different parts of the network, leading Ay to formulate the abstract mathematical problem to characterize the maximizers of the information divergence from an exponential family [5]. Both principles will be discussed in Section 5.1.

The case where the exponential family is the independence model $\mathcal{E}_1$ of a finite set of random variables $X_1, \ldots, X_n$ was treated by Ay and Knauf in [7]. In this case the information divergence of the joint distribution P of $X_1, \ldots, X_n$ is also called the *multiinformation* $I(X_1; \ldots; X_n) := D(P\|\mathcal{E}_1)$. Assume that X_i takes values in the finite set $\mathcal{X}_i$ of cardinality N_i, and assume that $N_n = \max_i N_i$. Using the chain rule for the entropy, Ay and Knauf observed that the multiinformation is bounded from above by $\sum_{i=1}^{n-1} \log(N_i)$. This upper bound is tight if and only if

$$N_n \geq \sum_{A \subseteq \{1,\ldots,n\}, S \neq \emptyset} (-1)^{|S|-1} \gcd\left(\{N_i : i \in S\}\right), \tag{1.1}$$

where gcd denotes the greatest common divisor. If this inequality holds, then there is an easy description of the set of global maximizers. This covers, in particular, the homogeneous case $N_1 = N_2 = \cdots = N_n$ and the case $n = 2$. The smallest numbers that violate this inequality are $n = 3$, $N_1 = N_2 = 3$ and $N_3 = 2$. Example 3.44 in Section 3.7 of this thesis computes the global maximum of the multiinformation in this case, which was unknown before.

In [51] František Matúš generalized the methods of [7] and derived inequalities that yield upper bounds on the information divergence for arbitrary hierarchical models (see Section 2.4 for the definition of a hierarchical model used in this thesis). If these bounds are tight, then global maximizers can be found by solving combinatorial problems. For example, the global maximum of the information divergence from the pair interaction model of four random variables of cardinality $N_4 = N_3 = N_2 = N_1$ is related to the existence of two orthogonal Latin squares of size N_1. Since there are no two orthogonal Latin squares of size two, the global maximum of the information divergence was previously unknown in the case $N_1 = 2$. This global maximum is computed in Example 3.42 in Section 3.6.

The problem simplifies if it is possible to compute the rI-projection map $P \mapsto P_{\mathcal{E}}$ in closed form. This is the case if the exponential family is convex, and this example is treated by Matúš and Ay in [52]. They describe the set of local and global maximizers and find criteria when the global maximizers are isolated. A special class of convex exponential families, the partition exponential families, plays an important role in Chapters 4 and 5. Other exponential families where a closed form of the rI-projection map is known are the binary i.i.d. sequences and binomial models. In [49] Matúš considers these one-dimensional examples and computes the global maximizers. The results were recently generalized by Juríček to multinomial models [41]. Section 4.4 revisits the binary i.i.d models and the binomial models and finds the local maximizers.

In 2007, Matúš computed the full first order optimality conditions of $D_{\mathcal{E}}$ [50], generalizing results of [5]. His analysis shows that all local maximizers P of $D_{\mathcal{E}}$ satisfy the projection property, that is, P equals the truncation of its rI-projection $P_{\mathcal{E}}$ onto $\mathcal{E}$ to its support:

$$P(x) = \begin{cases} \frac{1}{P_{\mathcal{E}}(\mathrm{supp}(P))} P_{\mathcal{E}}(x), & \text{if } P(x) > 0, \\ 0, & \text{else.} \end{cases}$$

This projection property will be discussed in Section 3.2; it is the basis for the theory developed in Chapter 3.

Apart from the original motivation there are further reasons to study the maximization of the information divergence: In the case of an independence model of two random variables X, Y, the information divergence is a natural measure of statistical dependence of random variables, called the *mutual information* $I(X;Y)$ (this is a special case of the multiinformation mentioned above). In [76] Jana Zvárová presented axioms that such measures should satisfy; for example, such a measure should be normalized to take values between zero and one. The mutual information itself does not satisfy this constraint, but any upper bound of the form $I(X;Y) \leq g(X;Y)$ yields a normalized measure $i_g(X;Y) := I(X;Y)/g(X;Y)$ (where some care has to be taken if $g(X;Y) = 0$). The function g determines under which conditions $i_g(X;Y) = 1$. For the trivial choice $g(X;Y) = I(X;Y)$ the measure i_g is constant. Therefore, depending on the applications one has in mind, further axioms should be required on i_g, and such axioms lead to conditions on g. Zvárová suggests to use Shannon's inequality

$$I(X;Y) \leq \min\{H(X); H(Y)\},$$

where H is the Shannon entropy. With this choice $i_g(X;Y) = 1$ if and only if X is a function of Y or vice versa. The main problem discussed in this thesis corresponds to the case where g depends only on the ranges of the random variables X and Y.

Similarly, the information divergence from the interaction models (see Section 2.4) can be interpreted as a measure of complexity. One possible definition of complexity states that a composite system is complex if its behaviour cannot be understood by analyzing its subsystems independently. In [4] Ay proposed to formalize this idea by measuring the distance of the state of the system from some suitably defined product

state. In the special case that the distance measure is the information divergence and the state of the system is described by a probability distribution this leads to the multiinformation. Later, e.g. in [8, 42], the multiinformation was decomposed into terms corresponding to different interaction orders. Upper bounds on the information divergence can be used, as above, to define normalized measures of complexity. Information about the maximizing probability distributions is useful when studying under which conditions such normalized measures are maximal. Complexity measures will be discussed in Section 5.1.

Both the infomax and the IMI principle take into account constraints which may arise both from the structure of the network and from the environment in which the network is situated. Therefore, it is necessary to study the constrained maximization of the information divergence. Under appropriate constraints the problem is well-posed even for exponential families on infinite sets. Such constrained maximization problems also arise, for example, in information theory, where the capacity of a channel is the supremum of the multiinformation between input and output under arbitrary input distributions. The constrained problem is beyond the scope of this thesis. Apart from some remarks in Section 5.1 it will be left as an open problem for the future.

A third motivation for maximizing the information divergence is the search for small exponential families such that the maximum value of the information divergence is bounded by a constant D. The idea is that such exponential families can approximate arbitrary probability distributions well, up to a divergence of D. Yaroslav Bulatov proposed that such exponential families are useful in machine learning (personal communication). This problem is presented and studied in Section 5.2.

As with any optimization problem, a lot of insight can be gained from a first order analysis. The points at which all existing two-sided directional derivatives of $D_{\mathcal{E}}$ vanish satisfy the projection property; and they are interesting in their own right. For example, they play a special role in the empirical estimation of the information divergence. If the true probability distribution satisfies the projection property, then the asymptotic distribution is a generalized χ^2-distribution. Otherwise the empirical distribution is asymptotically normally distributed. This was observed for the case of independence models by Milan Studený in [68]. The general case is presented in Section 5.3.

New results and outline of this thesis. This thesis contributes mainly to the mathematical aspects of the maximization of $D_{\mathcal{E}}$. In Chapter 3 the first order optimality conditions are analyzed and yield a reformulation of the problem: Finding the maximizers of $D_{\mathcal{E}}$ is equivalent to finding the maximizers of a function $\overline{D}_{\mathcal{E}}$ that is defined on the boundary of a polytope $\mathbf{U}_{\mathcal{N}}$ in the normal space of the exponential family $\mathcal{E}$. The main theorem, relating the two functions $D_{\mathcal{E}}$ and $\overline{D}_{\mathcal{E}}$, is Theorem 3.28. This reformulation leads to both theoretical insights and new algorithms. Two examples demonstrate the effectivity of these algorithms. For these examples it is important that the optimization problem can be transformed into algebraic equations. This makes it possible to use computer algebra systems and to automatize a large part of the calcu-

lations. The relation to algebraic equations was known before, see Remark 3.46, but, as is often the case with algebraic equations, the running time of algebraic algorithms depends crucially on the form of the equations. From an algebraic viewpoint, the results of Chapter 3 lead to a significant reduction of the number of variables in these equations.

A second emphasis of this thesis is on examples. Chapter 4 discusses mainly two classes of exponential families: One-dimensional exponential families are easy to understand, since it is easy to parametrize the set of all one-dimensional exponential families. Moreover, the problem can be reduced to a collection of one-dimensional problems that involve the study of real functions on subintervals of the real line. This is true both for the maximization of $D_{\mathcal{E}}$ and the maximization of $\overline{D}_{\mathcal{E}}$. A second important class of examples are the partition exponential families. These exponential families have interesting properties with respect to the maximization of $D_{\mathcal{E}}$: The value of the maximum is comparatively low, in a certain sense.

The last chapter is dedicated to applications. The connection to neural learning, complexity and channel capacities can only be sketched in this thesis. For these applications it would be desirable to extend the results of Chapter 3 to account for constraints, an undertaking that is beyond the scope of this thesis. The second application, the search for small exponential families that can approximate arbitrary probability distributions up to a fixed information divergence, can be directly studied with the tools developed in this thesis. Section 5.2 contains some results on this problem. The results are of theoretical nature, but they yield ideas how to adjust machine learning algorithms, like the minimax algorithm. These ideas will be tested in practise in a future project.

This thesis is organized as follows: Chapter 2 reviews the basic properties of exponential families and the information divergence. Most results of that chapter are rather well-known, but the presentation of some of the results, focusing on the implicit description in Theorem 2.21, is novel. The chapter contains two sections on algebraic exponential families and hierarchical models.

The heart of this thesis is contained in Chapter 3. After a review on the first order optimality conditions of $D_{\mathcal{E}}$ and related results in Section 3.1, a new interpretation of these conditions is given in Section 3.2. This is used in Section 3.3 to show that the global maximizers of $D_{\mathcal{E}}$ are in bijection with the global maximizers of $\overline{D}_{\mathcal{E}}$. In Section 3.4 the first order optimality conditions of $\overline{D}_{\mathcal{E}}$ are computed. Section 3.5 shows that the local maximizers and the critical points of $D_{\mathcal{E}}$ can also be found by studying the local maximizers and critical points of $\overline{D}_{\mathcal{E}}$. Section 3.6 discusses how the critical equations can be solved systematically, using the help of computer algebra systems in the case where $\mathcal{E}$ is algebraic. Section 3.7 proposes a different algorithm, which uses the projection property of local maximizers of $D_{\mathcal{E}}$ more directly. Sections 3.6 and 3.7 contain detailed discussions of two examples to which the proposed algorithms have been applied.

Chapter 4 is dedicated to examples. In addition to results for concrete exponential families it also presents useful general methods to treat the problem. Section 4.1

studies low-dimensional exponential families, with an emphasis on the one-dimensional case. The partition models, introduced in Section 4.2, are convex exponential families and have a low global maximum value of $D_{\mathcal{E}}$. The results of Chapter 3 imply that $\max D_{\mathcal{E}} \geq \log(2)$, unless $\mathcal{E}$ contains all strictly positive probability measures. The exponential families where this bound is achieved are found in Section 4.3; it turns out that the minimal such exponential families are partition models. Partition models also appear in the study of symmetries. The set of all probability measures that are symmetric under the action of a given group of permutations of the ground set $\mathcal{X}$ is a partition model, and exponential families that consist of symmetric probability measures are subfamilies of partition models. Section 4.4 illustrates how to exploit this fact using the example of the binomial models and binary i.i.d. models.

Chapter 5 is dedicated to the possible applications: Section 5.1 explains the connection to learning theory and complexity measures and comments on the importance of generalizing the results from this thesis to the constrained maximization of the information divergence. Section 5.2 studies small exponential families with bounded maximum information divergence. Many such exponential families that are optimal in a certain sense are partition models. The asymptotic behaviour of the empirical information divergence is analyzed in Section 5.3. It is shown that the type of the asymptotic distribution changes if the underlying distribution satisfies the projection property.

Notation and conventions. In this work only probability spaces over finite sets are discussed (with one exception: Section 5.3 discusses a countable family of independent identically distributed random variables with values in a finite set $\mathcal{X}$). Therefore, technical details concerning, for example, null sets or the sigma algebra, are ignored. The sigma algebra on a finite set is always the full power set.

The cardinality of a set $\mathcal{X}$ is denoted by $|\mathcal{X}|$. For any natural number $n \in \mathbb{N}$ the set $\{1, \ldots, n\}$ is denoted by $[n]$. Let $\mathcal{X}$ be a finite set of cardinality $|\mathcal{X}| > 1$. The set of all functions on $\mathcal{X}$ with values in a set V is $V^{\mathcal{X}}$. The set of real numbers is denoted by $\mathbb{R}$, the set of nonnegative real numbers by $\mathbb{R}_{\geq}$. The set of complex numbers is $\mathbb{C}$, and $\mathbb{C}^{\times}$ is the multiplicative group of the field $\mathbb{C}$. The set of integers is denoted by $\mathbb{Z}$, and the set of natural numbers is $\mathbb{N}$. The floor function $\lfloor \cdot \rfloor$ and the ceiling function $\lceil \cdot \rceil$ are defined for all $a \in \mathbb{R}$ via

$$\lfloor a \rfloor = \max\{z \in \mathbb{Z} : z \leq a\} \qquad \text{and} \qquad \lceil a \rceil = \min\{z \in \mathbb{Z} : z \geq a\}.$$

A real function $\mu \in \mathbb{R}^{\mathcal{X}}$ is also called a *(signed) measure*. $\mathbb{R}^{\mathcal{X}}$ is a vector space with a basis $\{\delta_x\}_{x \in \mathcal{X}}$ given by the *point measures*

$$\delta_x(y) = \begin{cases} 1, & \text{if } x = y, \\ 0, & \text{else.} \end{cases}$$

The components of $\mu \in \mathbb{R}^{\mathcal{X}}$ are written as either $\mu(x)$ or μ_x. Real functions $f, g \in \mathbb{R}^{\mathcal{X}}$ can be multiplied pointwise, such that $(fg)(x) = f(x)g(x)$. Furthermore, $\mathbb{R}^{\mathcal{X}}$ has a

canonical scalar product

$$\langle u, v\rangle = \sum_{x\in\mathcal{X}} u(x)v(x).$$

This scalar product defines orthogonal complements $V^\perp$ of subsets $V \subseteq \mathbb{R}^{\mathcal{X}}$.

If $\mathcal{Y} \subseteq \mathcal{X}$, then $\mu(\mathcal{Y}) := \sum_{x\in\mathcal{Y}} \mu(x)$. A measure $\mu \neq 0$ is *positive* if all its components are nonnegative, i.e. $\mu(x) \in \mathbb{R}_{\geq}$ for all $x \in \mathcal{X}$. It is *strictly positive* if all components are positive. Arbitrary nonzero measures $\mu \neq 0$ can be uniquely decomposed $\mu = \mu^+ - \mu^-$ as a difference of two positive measures μ^+, μ^- such that $\mu^+\mu^- = 0$. A measure μ is *normalized* if $\mu(\mathcal{X}) = 1$. A normalized positive measure is a *probability measure*. The set of all probability measures on $\mathcal{X}$ is denoted by $\mathbf{P}(\mathcal{X})$. The set of all strictly positive probability measures is denoted by $\mathbf{P}(\mathcal{X})^\circ$.

The following convention from probability theory is useful: If $u \in \mathbb{R}^{\mathcal{X}}$, then equalities and inequalities like $u = 0$, $u \neq 0$, $u < 0$, and so on, are interpreted as their corresponding solution sets when they appear as the argument of a measure. For example, $\mu(u \neq 0) = \sum_{x:u(x)\neq 0} \mu(x)$.

Any map $\phi : \mathcal{X} \to \mathcal{Z}$ induces a natural map $\phi^* : \mathbb{R}^{\mathcal{Z}} \to \mathbb{R}^{\mathcal{X}}$ by pullback, such that $\phi^* v(x) = v(\phi(x))$. If ϕ is a surjection, then ϕ^* is an injection, so $\mathbb{R}^{\mathcal{Z}}$ can be considered as a subset of $\mathbb{R}^{\mathcal{X}}$. There is also a pushforward ϕ_*, defined by $\phi_* v(A) = v(\phi^{-1}(A))$. For example, for any subset $\mathcal{Y} \subseteq \mathcal{X}$ the set of functions $\mathbb{R}^{\mathcal{Y}}$ on $\mathcal{Y}$ can be seen as a subset of $\mathbb{R}^{\mathcal{X}}$, with $v(x) = 0$ for all $v \in \mathbb{R}^{\mathcal{Y}}$ and $x \in \mathcal{X} \setminus \mathcal{Y}$.

The *support* of a measure μ on $\mathcal{X}$ is defined as $\operatorname{supp}(\mu) = \{x \in \mathcal{X} : \mu(x) \neq 0\}$. If $\operatorname{supp}(\mu) = \mathcal{X}$, then μ has *full support*. The *restriction* of μ to a subset $\mathcal{Y} \subseteq \mathcal{X}$ is the measure $\mu|_{\mathcal{Y}} \in \mathbb{R}^{\mathcal{Y}}$ given by $\mu|_{\mathcal{Y}}(x) = \mu(x)$ for all $x \in \mathcal{Y}$. If a positive measure μ on $\mathcal{X}$ satisfies $\mu(\mathcal{Y}) > 0$, then the *truncation* of μ to $\mathcal{Y}$ is the probability measure $\mu^{\mathcal{Y}} = \frac{1}{\mu(\mathcal{Y})}\mu|_{\mathcal{Y}} \in \mathbf{P}(\mathcal{Y}) \subseteq \mathbf{P}(\mathcal{X})$ satisfying

$$\mu^{\mathcal{Y}}(x) = \begin{cases} \frac{1}{\mu(\mathcal{Y})}\mu(x), & \text{if } x \in \mathcal{Y}, \\ 0, & \text{else.} \end{cases}$$

If $P \in \mathbf{P}(\mathcal{X})$ satisfies $P(\mathcal{Y}) > 0$, then the truncation $P^{\mathcal{Y}}$ is also written $P(\cdot|\mathcal{Y})$ and called the *conditional probability* of P given $\mathcal{Y}$.

The constant function $(1, \ldots, 1) \in \mathbb{R}^{\mathcal{X}}$ is denoted by $\mathbf{1}$. It is also called the *uniform measure* on $\mathcal{X}$. The *uniform distribution* is $\frac{1}{|\mathcal{X}|}\mathbf{1} = \mathbf{1}^{\mathcal{X}}$. More generally, the *characteristic function* $\mathbf{1}_{\mathcal{Y}}$ of a subset $\mathcal{Y} \subseteq \mathcal{X}$ is

$$\mathbf{1}_{\mathcal{Y}}(x) = \begin{cases} 1, & \text{if } x \in \mathcal{Y}, \\ 0, & \text{else.} \end{cases}$$

All logarithms are natural logarithms. This convention plays a role only in few parts of this thesis: In Section 3.6 the fact that different branches of the logarithm differ by an integral multiple of $2\pi\mathfrak{i}$ is used. Furthermore, the approximate numerical values given in Sections 3.6 and 3.7 as well as the scale for the various figures is given in units corresponding to the natural logarithm (such units are sometimes called *nits*, as opposed to *bits*).

2. Exponential families

The purpose of this chapter is to define exponential families and to collect some basic properties and tools. An early geometric treatment on exponential families is [10], a more detailed textbook is [15]. Exponential families also play an important role in information geometry, see [3]. The most general results about closures and boundaries of exponential families can be found in a series of papers by Csiszár and Matúš, see [21]; earlier results are due to Chentsov [16]. The first three chapters of [20] are a short introduction to the relation between exponential families and the information divergence.

Some of the properties presented in this chapter are well-known. Unfortunately, there exists no reference including all the necessary results. Furthermore, the notation varies from one author to the next, and sometimes also the definitions. Finally, in this thesis only the case of a finite ground set will be needed, which leads to a considerable simplification of many results. For these reasons most proofs are included in a generality adapted to the applications in mind.

Section 2.1 defines exponential families and relates them to the information divergence. The proofs of the results on the closure and the boundary in Section 2.2 are organized such as to highlight the usefulness of the implicit representation, given in Theorem 2.21. This result seems to be new in this form and was published in [61]. The implicit representation generalizes the fact that algebraic exponential families have an implicit description by polynomials, a result due to Geiger, Meek and Sturmfels [31]. This algebraic result is presented in Section 2.3 together with an overview of other algebraic notions that are used in this thesis. The most important class of algebraic exponential families are the hierarchical models, discussed in Section 2.4.

2.1. Exponential families, the convex support and the moment map

Throughout this chapter let $\mathcal{X}$ be a finite set of cardinality N.

Definition 2.1. Let ν be a strictly positive measure on $\mathcal{X}$. Let $\mathcal{T} \subseteq \mathbb{R}^{\mathcal{X}}/\mathbb{R}\mathbf{1}$ be a linear subspace of the vector space of functions on $\mathcal{X}$ modulo the constant functions, and write $\tilde{\mathcal{T}} = \{f \in \mathbb{R}^{\mathcal{X}} : f + \mathbb{R}\mathbf{1} \in \mathcal{T}\}$. The *exponential family* $\mathcal{E}_{\nu,\mathcal{T}}$ with *reference measure* ν and *tangent space* $\mathcal{T}$ consists of all probability measures on $\mathcal{X}$ of the form

$$P_\theta(x) = \frac{\nu_x}{Z_\theta} \exp\left(\theta(x)\right), \tag{2.1}$$

where $\theta \in \tilde{\mathcal{T}}$ and Z_θ ensures normalization. $\tilde{\mathcal{T}}$ is called the *extended tangent space* of $\mathcal{E}$. The orthogonal complement $\mathcal{N} = \tilde{\mathcal{T}}^\perp$ (with respect to the canonical scalar product on $\mathbb{R}^\mathcal{X}$) is called the *normal space* of $\mathcal{E}_{\nu,\mathcal{T}}$.

The topological closure of $\mathcal{E}_{\nu,\mathcal{T}}$ is denoted by $\overline{\mathcal{E}_{\nu,\mathcal{T}}}$. The *boundary* of $\mathcal{E}_{\nu,\mathcal{T}}$ is $\overline{\mathcal{E}_{\nu,\mathcal{T}}} \setminus \mathcal{E}_{\nu,\mathcal{T}}$.

Remark 2.2. The reason that $\mathcal{T}$ is defined as a subspace of $\mathbb{R}^\mathcal{X}/\mathbb{R}\mathbf{1}$ is that the effect of $\mathbf{1}$ itself is cancelled by the normalization condition. To be precise, if $\theta - \theta' \in \mathbb{R}\mathbf{1}$, then $P_\theta = P_{\theta'}$, and conversely. Therefore, one can always assume $Z_\theta = 1$ by choosing θ correspondingly.

Remark 2.3. The restriction that ν is strictly positive can be relaxed. However, if $\nu_x = 0$ for some $x \in \mathcal{X}$, then $P(x) = 0$ for all $P \in \overline{\mathcal{E}_{\nu,\mathcal{T}}}$, so for most considerations x can be removed from $\mathcal{X}$. Furthermore, in this case the maximum value of the information divergence, studied in the next chapter, is always infinite, see Proposition 2.14.

Sometimes it is convenient to require that ν be a probability measure, but in other cases it simplifies the notation to allow arbitrary reference measures. For example, in the case of a uniform reference measure a factor of $\frac{1}{|\mathcal{X}|}$ is needed to normalize $\mathbf{1}$.

Remark 2.4. The nomenclature used in this thesis is slightly unorthodox. Usually, an exponential family $\mathcal{E}$ is defined using an explicit parametrization of $\mathcal{E}$ with the help of a generating set of $\tilde{\mathcal{T}}$ as in (2.2) below. The notions of (extended) tangent space and normal space are introduced here in order to allow a formulation of the results that is invariant from the choice of a parametrization. The nomenclature is justified by the following considerations:

The differential geometric tangent space of $\mathcal{E}$ in a point $P \in \mathcal{E}$ equals the image of $\tilde{\mathcal{T}}$ under the differential of $\theta \mapsto P_\theta$. Computing the differential of (2.1) (cf. Section 5.3) shows

$$\mathcal{T}_P = \left\{ fP : f \in \tilde{\mathcal{T}}, \sum_x f(x)P(x) = 0 \right\}.$$

The natural map $\mathcal{T} \to \mathcal{T}_P$, $f + \mathbb{R}\mathbf{1} \mapsto (f - \sum_x f(x)P(x))P$ is an isomorphism, justifying the name *tangent space* for $\mathcal{T}$. In particular, the dimension of $\mathcal{E}$ (as a manifold) equals $\dim \mathcal{T} = \dim \tilde{\mathcal{T}} - 1$.

If $\mathcal{X}$ is infinite (and if the notion of an exponential family is suitably generalized), then there are different notions of closure for an exponential family, see [21]. In the finite case, they all agree. A similar remark applies to the setting of quantum statistics, see [71]. Since the maximal value of the function $D_\mathcal{E}$ that will be studied in the next chapter is usually infinite if $\mathcal{X}$ is infinite, only the finite case is considered here. Possible generalizations to the quantum case are beyond the scope of this thesis.

$\mathcal{E}_{\nu,\mathcal{T}}$ can be parametrized by choosing functions $\{a_i\}_{i=1}^h \subset \mathbb{R}^\mathcal{X}$ such that the $a_i + \mathbb{R}\mathbf{1}$ generate $\mathcal{T}$. It is convenient to arrange this generating set as rows in a matrix $A \in \mathbb{R}^{h \times \mathcal{X}}$ such that $A_{i,x} = a_i(x)$. Then $\mathcal{E}_{\nu,\mathcal{T}}$ consists of all probability measures on $\mathcal{X}$ of the form

$$P_\vartheta(x) = \frac{\nu_x}{Z_\vartheta} \exp\left(\sum_{i=1}^h \vartheta_i A_{i,x} \right), \tag{2.2}$$

where $\vartheta \in \mathbb{R}^h$. Alternatively, the *monomial parametrization*

$$P_\xi(x) = \frac{\nu_x}{Z_\xi} \prod_{i=1}^{h} \xi_i^{A_{i,x}} \tag{2.3}$$

can be used, where $\xi \in \mathbb{R}^h$, $\xi_i > 0$ for all i (the name comes from the fact that if all entries $A_{i,x}$ are nonnegative integers, then the mapping $\xi \mapsto P_\xi(x)$ is indeed an algebraic monomial for all $x \in \mathcal{X}$, cf. Section 2.3). Conversely, using either (2.2) or (2.3), an exponential family $\mathcal{E}_{\nu,A}$ can be associated to any reference measure ν and to any matrix $A \in \mathbb{R}^{h\times\mathcal{X}}$.

Definition 2.5. Any matrix $A \in \mathbb{R}^{h\times\mathcal{X}}$ such that $\mathcal{E}_{\nu,\mathcal{T}} = \mathcal{E}_{\nu,A}$ is called a *sufficient statistics* of $\mathcal{E}_{\nu,\mathcal{T}}$.

For an interpretation of the name "sufficient statistics" and its meaning outside of the theory of exponential family see [10] and [17].

The following lemma follows easily from the definitions:

Lemma 2.6. *Let $A, A' \in \mathbb{R}^{h\times\mathcal{X}}$ and let $\nu, \nu' \in \mathbb{R}^{\mathcal{X}}$ be strictly positive measures. Then $\mathcal{E}_{\nu,A} = \mathcal{E}_{\nu',A'}$ if and only if the following two conditions are satisfied:*

- *The probability measure $\frac{1}{\sum_{x\in\mathcal{X}} \nu_x}\nu$ lies in $\mathcal{E}_{\nu',A'}$.*
- *The row space of A equals the row space of A' modulo $\mathbf{1}$.*

Definition 2.7. Let $\mathcal{E}_{\nu,A}$ be an exponential family. The linear map

$$\pi_A : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^h, v \mapsto Av$$

corresponding to the sufficient statistics A is called the *moment map*. The image of $\mathbf{P}(\mathcal{X})$ under the moment map is a polytope $\mathbf{M}_A$, called the *convex support* of $\mathcal{E}_{\nu,A}$.

Every $x \in \mathcal{X}$ can be identified with the point $A_x \in \mathbf{M}_A$ given by the corresponding column of A. The polytope $\mathbf{M}_A$ equals the convex hull of these points. Therefore, every vertex of $\mathbf{M}_A$ is of the form A_x for some $x \in \mathcal{X}$; but in general not every point A_x needs to be a vertex. See Appendix A.1 for a summary of basic facts about polytopes.

The polytope $\mathbf{M}_A$ is, up to affine equivalence, independent of the choice of A (see Remark 2.23 for an invariant characterization of the face lattice of $\mathbf{M}_A$):

Lemma 2.8. *Let $A \in \mathbb{R}^{h\times\mathcal{X}}$ and $A' \in \mathbb{R}^{h'\times\mathcal{X}}$ be two matrices. If $\mathcal{E}_{\nu,A} = \mathcal{E}_{\nu',A'}$ for some reference measures ν, ν', then there are linear maps $B : \mathbb{R}^h \to \mathbb{R}^{h'}$, $C : \mathbb{R}^{h'} \to \mathbb{R}^h$ and vectors $b \in \mathbb{R}^{h'}$, $c \in \mathbb{R}^h$ such that $b + BA_x = A'_x$ and $c + CA'_x = A_x$ for all $x \in \mathcal{X}$. In particular, $\mathbf{M}_A$ and $\mathbf{M}_{A'}$ are affinely equivalent. Conversely, if such affine maps exist, then $\mathcal{E}_{\nu,A} = \mathcal{E}_{\nu,A'}$ for all reference measures ν.*

Proof. By Lemma 2.6 every row a_i of A can be written as $a_i = \sum_{j=1}^{h'} b_{i,j}a'_j + b_i\mathbf{1}$. Let B be the matrix $(b_{i,j})_{i,j}$ and let b be the vector $(b_i)_i$; then $A_x = BA'_x + b$ for all $x \in \mathcal{X}$. The matrix C and the vector c are defined similarly, by exchanging the roles of A and A'. The last statement follows from Lemma 2.6. □

For some applications it is advantageous to choose A such that the constant function $\mathbf{1} \in \mathbb{R}^{\mathcal{X}}$ is contained in the row space of A. In this case $\tilde{\mathcal{T}}$ equals the row space of A, the normal space is $\mathcal{N} = \ker A$, and $\dim(\mathcal{E}_{\nu,\mathcal{T}}) = \operatorname{rank}(A) - 1$. On the other hand, one may choose the rows $\{a_i\}_{i=1}^h$ such that the row space of A is a vector space complement of $\mathbb{R}\mathbf{1}$ in $\tilde{\mathcal{T}}$. This reduces the number of parameters, such that $\operatorname{rank}(A) = \dim(\mathcal{E}_{\nu,A})$.

Lemma 2.9. *Let $A \in \mathbb{R}^{h\times\mathcal{X}}$. The following statements are equivalent:*

(*i*) *The row space of A contains $\mathbf{1}$.*

(*ii*) $\mathcal{N} = \ker A = \ker \pi_A$.

(*iii*) *Every $u \in \ker A$ satisfies $\sum_{x\in\mathcal{X}} u(x) = 0$.*

(*iv*) *The polytope $\mathbf{M}_A$ lies in a hyperplane in $\mathbb{R}^h$ that does not contain the origin.*

(*v*) *There is a dual vector $\ell \in (\mathbb{R}^h)^*$ such that $\ell(p) = 1$ for all $p \in \mathbf{M}_A$.*

Proof. The fact that the first two statements are equivalent was already remarked before the statement of the lemma. $\ker A$ is the orthogonal complement of the row space of A. (*i*) says that the constant vector $\mathbf{1}$ is orthogonal to $\ker A$; so the first two statements are equivalent to (*iii*). The equivalence of the last two statements follows directly from the definition of a hyperplane.

If (*v*) holds, then $0 = \ell(Au) = \sum_{x\in\mathcal{X}} \ell(A_x)u(x) = \sum_{x\in\mathcal{X}} u(x)$, so (*iii*) holds. Conversely, if (*i*) holds, then write $\mathbf{1}$ as a linear combination $\sum_{i=1}^h \ell_i a_i$ of the rows a_i of A. Then $\ell(\vartheta) := \sum_{i=1}^h \ell_i\vartheta_i$ defines a dual vector $\ell \in (\mathbb{R}^h)^*$ satisfying $\ell(A_x) = 1$ for all $x \in \mathcal{X}$. This implies (*v*). □

Definition 2.10. An *exponential subfamily* of an exponential family $\mathcal{E}$ is an exponential family $\mathcal{E}'$ such that $\mathcal{E}' \subseteq \mathcal{E}$.

Lemma 2.11. *If $\mathcal{E}'$ is an exponential subfamily of $\mathcal{E}$, then any reference measure of $\mathcal{E}'$ is a reference measure of $\mathcal{E}$, the tangent space of $\mathcal{E}'$ is a linear subspace of the tangent space of $\mathcal{E}$, and the convex support of $\mathcal{E}'$ is an affine image of the convex support of $\mathcal{E}$.*

Proof. If ν is a reference measure of $\mathcal{E}'$, then $\frac{1}{\nu(\mathcal{X})}\nu \in \mathcal{E}' \subseteq \mathcal{E}$. This implies the first statement. The tangent spaces $\mathcal{T}'$ and $\mathcal{T}$ of $\mathcal{E}'$ and $\mathcal{E}$ satisfy

$$\mathcal{T}' = \left\{ \left(\log \frac{P(x)}{Q(x)} \right)_{x\in\mathcal{X}} : P, Q \in \mathcal{E}' \right\} \subseteq \left\{ \left(\log \frac{P(x)}{Q(x)} \right)_{x\in\mathcal{X}} : P, Q \in \mathcal{E} \right\} = \mathcal{T}.$$

Let $A' \in \mathbb{R}^{h'\times\mathcal{X}}$ and $A \in \mathbb{R}^{h\times\mathcal{X}}$ be sufficient statistics of $\mathcal{E}'$ and $\mathcal{E}$. Then any row a'_i of A' is a linear combination $a'_i = b_i + \sum_j B_{i,j}a_j$ of $\mathbf{1}$ and the rows a_j of A. Therefore $\mathbf{M}_{A'} = b + B\mathbf{M}_A$, with $b = (b_i)_i \in \mathbb{R}^{h'}$ and $B = (B_{i,j})_{i,j} \in \mathbb{R}^{h'\times h}$. □

The fibres of the restriction of π_A to $\mathbf{P}(\mathcal{X})$ are themselves important statistical models.

Definition 2.12. A *linear family* on a set $\mathcal{X}$ is the intersection of $\mathbf{P}(\mathcal{X})$ with an affine subspace of $\mathbb{R}^{\mathcal{X}}$. Let $P \in \mathbf{P}(\mathcal{X})$. If $\mathcal{E}$ is an exponential family on $\mathcal{X}$ with normal space $\mathcal{N}$, then write $\mathcal{N}_P$ for the linear family

$$\mathcal{N}_P := \{Q \in \mathbf{P}(\mathcal{X}) : P - Q \in \mathcal{N}\}.$$

Theorem 2.16 will show that there is a deep relation between exponential families and their corresponding linear families. This relation is best understood with regard to the information divergence:

Definition 2.13. The *information divergence* of two positive measures μ, ν on $\mathcal{X}$ is defined as

$$D(\mu\|\nu) = \sum_{x\in\mathcal{X}} \mu(x) \log\left(\frac{\mu(x)}{\nu(x)}\right),$$

with the convention that $0\log 0 = 0\log(0/0) = 0$. If there exists $x \in \mathcal{X}$ such that $\nu(x) = 0$ and $\mu(x) \neq 0$, then $D(\mu\|\nu) = +\infty$. The information divergence of a positive measure μ from an exponential family $\mathcal{E}$ is

$$D_{\mathcal{E}}(\mu) := D(\mu\|\mathcal{E}) = \inf_{Q\in\mathcal{E}} D(\mu\|Q).$$

The information divergence is also known under the name *Kullback-Leibler divergence* or *relative entropy*. It is most commonly used when μ and ν are probability measures. It has the following properties:

Proposition 2.14. *Let μ and ν be two positive measures on $\mathcal{X}$. Then:*

(i) $D(a\mu\|b\nu) = aD(\mu\|\nu) + a\mu(\mathcal{X})\log(a/b)$ *for all* $a, b > 0$.

(ii) $D(\mu\|\nu) = \infty$ *if and only if* $\operatorname{supp}(\nu) \not\subseteq \operatorname{supp}(\mu)$.

(iii) The function $(\mu, \nu) \mapsto D(\mu\|\nu)$ is convex. For fixed ν it is strictly convex in μ.

If P and Q are two probability measures on $\mathcal{X}$, then:

(iv) $D(P\|Q) \geq 0$, *and* $D(P\|Q) = 0$ *if and only if* $P = Q$.

Proof. *(i)* and *(ii)* follow directly from the definition. *(iii)* and *(iv)* are consequences of the log sum inequality, which is a special case of Jensen's inequality, see Chapters 2.6 and 2.7 in [17]. □

The information divergence is not continuous (because $\lim_{p\to 0, q\to 0} \frac{p}{q}$ does not exist), but the following holds:

Lemma 2.15. *Let $\mathcal{E} \subseteq \mathbf{P}(\mathcal{X})^\circ$. The function $\mu \mapsto D_{\mathcal{E}}(\mu)$ is continuous.*

Proof. See [5, Lemma 4.2]. □

The following theorem sums up the main facts about exponential families and the information divergence:

Theorem 2.16. *Let $\mathcal{E}$ be an exponential family on $\mathcal{X}$ with normal space $\mathcal{N}$ and reference measure ν, and let $P \in \mathbf{P}(\mathcal{X})$. Then there exists a unique probability measure $P_\mathcal{E} \in \overline{\mathcal{E}} \cap \mathcal{N}_P$. Furthermore, $P_\mathcal{E}$ has the following properties:*

(*i*) *For all $Q \in \overline{\mathcal{E}}$ the* Pythagorean identity

$$D(P\|Q) = D(P\|P_\mathcal{E}) + D(P_\mathcal{E}\|Q). \tag{2.4}$$

holds. In particular, $D(P\|\mathcal{E}) = D(P\|P_\mathcal{E})$.

(*ii*) *$P_\mathcal{E}$ satisfies*

$$D(P\|\mathcal{E}) = D(P\|\nu) - D(P_\mathcal{E}\|\nu). \tag{2.5}$$

(*iii*) $D(P_\mathcal{E}\|\nu) = \min\{D(Q\|\nu) : Q \in \mathcal{N}_P\}$.

Proof. Corollary 3.1 of [20] proves existence and uniqueness of $P_\mathcal{E}$ as well as the Pythagorean identity (2.4) for all probability measures P and all probability measures $Q \in \overline{\mathcal{E}}$. Statement (*ii*) follows from (*i*) and Proposition 2.14 (*i*). If $Q \in \mathcal{N}_P$, then $\mathcal{N}_Q = \mathcal{N}_P$. Hence (*ii*) implies $D(Q\|\nu) = D(Q\|\mathcal{E}) + D(P_\mathcal{E}\|\nu) \geq D(P_\mathcal{E}\|\nu)$. □

Definition 2.17. The probability measure $P_\mathcal{E}$ in Theorem 2.16 is called the *(generalized) reverse information projection* or *rI-projection* of P onto $\mathcal{E}$.

Remark 2.18. The attribute "generalized" is often used to distinguish the case that $P_\mathcal{E}$ lies in the boundary $\overline{\mathcal{E}} \setminus \mathcal{E}$, see [22]. This distinction is important in statistical applications, since in this case there are no parameters ϑ such that $P_\mathcal{E} = P_\vartheta$, and this must be taken care of when computing $P_\mathcal{E}$. In this work this problem will play no big role, and the attribute "generalized" will be omitted.

Remark 2.19. In the important case that $\nu = \mathbf{1}$ the function $D(P\|\nu)$ equals minus the *(Shannon) entropy*:

$$H(P) := -\sum_{x\in\mathcal{X}} P(x)\log P(x) = -D(P\|\mathbf{1}).$$

Remark 2.20. In general there is no analytical formula for the rI-projection $P_\mathcal{E}$ of a probability measure P. There are, however, relatively fast algorithms to compute $P_\mathcal{E}$. Most of them are iterative algorithms, like *iterative scaling* (see [23] and references therein; see also [20]). An implementation for hierarchical models is given by cipi [67].

2.2. The closure of an exponential family

Recall that the boundary of an exponential family $\mathcal{E}$ is defined as $\overline{\mathcal{E}} \setminus \mathcal{E}$. Probability measures in the boundary are not reachable by the canonical parametrization (2.2). There are three possibilities to work around this problem: First, Theorem 2.16 implies that the restriction of the moment map π_A is a bijection $\overline{\mathcal{E}} \mapsto \mathbf{M}_A$. Therefore, $\overline{\mathcal{E}}$ can be parametrized with the help of the polytope $\mathbf{M}_A$. Unfortunately, there are no analytic

formulas for this parametrization in general, see Remark 2.20. Second, it is possible to choose an appropriate sufficient statistics and allow the parameters in the monomial parametrization (2.3) to become zero. This approach is discussed in [61]. The third possibility is to work with an implicit description of $\mathcal{E}$.

It is convenient to use the following notation: For any two functions $u, p \in \mathbb{R}^{\mathcal{X}}$ define

$$p^u := \prod_{x \in \mathcal{X}} p(x)^{u(x)},$$

whenever this product is well defined (e.g. when u and p are both non-negative). Here $0^0 = 1$ by convention. Any $u \in \mathbb{R}^{\mathcal{X}}$ can be decomposed uniquely into a positive part $u^+ \in \mathbb{R}^{\mathcal{X}}_{\geq}$ and a negative part $u^- \in \mathbb{R}^{\mathcal{X}}_{\geq}$ such that $u = u^+ - u^-$ and $\operatorname{supp}(u^+) \cap \operatorname{supp}(u^-) = \emptyset$.

Theorem 2.21. *Let $\mathcal{E}$ be an exponential family with normal space $\mathcal{N}$ and reference measure ν. A probability measure P on $\mathcal{X}$ belongs to $\overline{\mathcal{E}}$ if and only if P satisfies*

$$P^{u^+}\nu^{u^-} = P^{u^-}\nu^{u^+}, \qquad \text{for all } u \in \mathcal{N}. \tag{2.6}$$

The proof is based on a lemma and the following definition:

Definition 2.22. Let $\mathcal{V}$ be a linear subspace of $\mathbb{R}^{\mathcal{X}}$. A subset $\mathcal{Y} \subseteq \mathcal{X}$ is *facial* (with respect to $\mathcal{V}$) if there exists a vector $\theta_0 \in \mathcal{V}$ and constants $a < b$ such that

$$\theta_0(y) = a \quad \text{for all } y \in \mathcal{Y}, \qquad\qquad \theta_0(z) \geq b \quad \text{for all } z \notin \mathcal{Y}. \tag{2.7}$$

$\mathcal{Y}$ is *facial* with respect to an exponential family $\mathcal{E}$ if $\mathcal{Y}$ is facial with respect to the extended tangent space $\tilde{\mathcal{T}}$ of $\mathcal{E}$. In this case one may assume $a = 0$ and $b = 1$ since $\mathbf{1} \in \tilde{\mathcal{T}}$.

Remark 2.23. Let $A \in \mathbb{R}^{h \times \mathcal{X}}$ be a sufficient statistics of an exponential family $\mathcal{E}$. A subset $\mathcal{Y} \subseteq \mathcal{X}$ is facial (with respect to $\mathcal{E}$) if there exists a vector $\tau \in \mathbb{R}^h$ and constants $a < b$ such that

$$\sum_i \tau_i A_{i,y} = a \quad \text{for all } y \in \mathcal{Y}, \qquad\qquad \sum_i \tau_i A_{i,z} \geq b \quad \text{for all } z \notin \mathcal{Y}.$$

This can be rephrased as follows: A set $\mathcal{Y} \subseteq \mathcal{X}$ corresponds to a collection of points in $\mathbf{M}_A$. Let $\mathbf{F}$ be the smallest face of the polytope $\mathbf{M}_A$ that contains all these points. Then $\mathcal{Y}$ is facial if and only if $\mathcal{Y} = \{x \in \mathcal{X} : A_x \in \mathbf{F}\}$. If A satisfies the statements of Lemma 2.9, then one may additionally require $a = 0$ and $b = 1$.

Lemma 2.24. *If P satisfies* (2.6)*, then* $\operatorname{supp}(P)$ *is facial with respect to $\mathcal{E}$.*

Proof. If $\mathcal{Y} := \operatorname{supp}(P)$ is not facial, then the intersection of the extended tangent space $\tilde{\mathcal{T}}$ and the set $\mathbf{B} := \{\theta \in \mathbb{R}^{\mathcal{X}} : \theta(x) = 0 \text{ for } x \in \mathcal{Y}, \theta(x) \geq 1 \text{ for } x \notin \mathcal{Y}\}$ is empty. Consider the projection $\theta \mapsto \theta + \tilde{\mathcal{T}}$ along $\tilde{\mathcal{T}}$. The image of $\tilde{\mathcal{T}}$ is the origin $0 \in \mathbb{R}^{\mathcal{X}}/\tilde{\mathcal{T}}$, and the image of $\mathbf{B}$ is a polyhedral set $\tilde{\mathbf{B}}$. By assumption $0 \notin \tilde{\mathbf{B}}$. By

Theorem A.3 there exists a linear form $\ell : \mathbb{R}^{\mathcal{X}}/\tilde{\mathcal{T}} \to \mathbb{R}$ that strictly separates $\tilde{\mathbf{B}}$ and 0. Composition with the projection yields a vector $u \in \mathbb{R}^{\mathcal{X}}$ and $c \in \mathbb{R}$ such that $\sum_{x\in\mathcal{X}} u(x)\theta(x) < c < \sum_{x\in\mathcal{X}} u(x)\theta'(x)$ for all $\theta \in \tilde{\mathcal{T}}$ and all $\theta' \in \mathbf{B}$. If $\theta \in \mathbb{R}^{\mathcal{X}}$ is not orthogonal to u, then $\lim_{\lambda\to\pm\infty} \sum_{x\in\mathcal{X}} \theta(x)\lambda u(x) \to \pm\infty$. Hence $u \in \mathcal{N}$ and $c > 0$. If $u(y) < 0$ for some $y \in \mathcal{X}$, then $\lim_{\lambda\to\infty} \left(\sum_{x\in\mathcal{X}\setminus(\mathcal{Y}\cup\{y\})} u(x) + u(y)\lambda\right) = -\infty < c$. Hence $u(x) > 0$ for all $x \notin \mathcal{Y}$. But this contradicts (2.6), since the left hand side of (2.6) vanishes, while the right hand side is greater than zero. □

Proof of Theorem 2.21. For any probability measure P on $\mathcal{X}$ let $P^{\mathbf{1}}(x) := \frac{1}{Z}\frac{P(x)}{\nu_x}$, where $Z := \sum_{x\in\mathcal{X}} \frac{P(x)}{\nu_x}$. Then $P \in \overline{\mathcal{E}}$ if and only if $P^{\mathbf{1}} \in \overline{\mathcal{E}_{\mathbf{1},A}}$; and P satisfies (2.6) if and only if $P^{\mathbf{1}}$ satisfies the same equations with ν replaced by $\mathbf{1}$. Therefore it suffices to consider the case $\nu = \mathbf{1}$.

Equations (2.6) hold on $\overline{\mathcal{E}}$: By continuity, it is enough to check this on $\mathcal{E}$. On $\mathcal{E}$ this is equivalent to $P^u = 1$ for all $u \in \mathcal{N}$, so the statement follows from the calculation

$$P_\theta^u = \left(\frac{1}{Z_\theta}\right)^{\sum_x u(x)} \exp\left(\sum_{x\in\mathcal{X}} \theta(x)u(x)\right) = 1,$$

using $\sum_x u(x) = 0 = \sum_x \theta(x)u(x)$.

For the other direction, suppose that P satisfies (2.6). Let $\mathcal{Y} := \operatorname{supp}(P)$, and define $l \in \mathbb{R}^{\mathcal{Y}}$ via $l(x) = \log(P(x))$ for all $x \in \mathcal{Y}$. Suppose that $l \neq \theta|_{\mathcal{Y}}$ for all $\theta \in \tilde{\mathcal{T}}$. Then there exists $v \in \mathbb{R}^{\mathcal{Y}} \subseteq \mathbb{R}^{\mathcal{X}}$ such that $\sum_{x\in\mathcal{Y}} l(x)v(x) \neq 0$ and $0 = \sum_{x\in\mathcal{Y}} v(x)\theta(x) = \sum_{x\in\mathcal{X}} v(x)\theta(x)$ for all $\theta \in \tilde{\mathcal{T}}$, whence $v \in \mathcal{N}$. But then

$$0 \neq \prod_{x\in\mathcal{Y}} P(x)^{v^+(x)} - \prod_{x\in\mathcal{Y}} P(x)^{v^-(x)} = \prod_{x\in\mathcal{X}} P(x)^{v^+(x)} - \prod_{x\in\mathcal{X}} P(x)^{v^-(x)},$$

in contradiction to the assumptions.

By Lemma 2.24 the set $\mathcal{Y}$ is facial. Choose $\theta_0 \in \tilde{\mathcal{T}}$ as in (2.7) with $a = 0$ and $b = 1$, and let $\theta_t := \theta + t\theta_0$, where $\theta \in \tilde{\mathcal{T}}$ satisfies $l = \theta|_{\mathcal{Y}}$. Then $\lim_{t\to-\infty} P_{\theta_t}(x) = 0$ for $x \notin \mathcal{Y}$, and $\lim_{t\to-\infty} P_{\theta_t}(x) = P(x)$ for $x \in \mathcal{Y}$, proving $P \in \overline{\mathcal{E}}$. □

The implicit characterization of $\overline{\mathcal{E}}$ in Theorem 2.21 consists of infinitely many equations. The following result shows that a finite subset of these equations suffices. Such a finite subset can be found using a circuit basis, which consists of precisely one circuit vector for every circuit, see Section A.2. A circuit determines its corresponding circuit vector up to a multiple. On the other hand, replacing $u \in \mathcal{N}$ by a nonzero multiple replaces (2.6) by an equation which is equivalent over the non-negative real numbers. This means that the systems of equations corresponding to different circuit bases C are all equivalent.

Theorem 2.25. *Let $\mathcal{E}$ be an exponential family, and let C be a circuit basis of its normal space $\mathcal{N}$. Then $\overline{\mathcal{E}}$ equals the set of all probability distributions P that satisfy*

$$P^{c^+}\nu^{c^-} = P^{c^-}\nu^{c^+} \qquad \text{for all } c \in C. \tag{2.8}$$

Proof. Again, one may assume that $\nu = \mathbf{1}$. It suffices to show: If $P \in \mathbf{P}(\mathcal{X})$ satisfies (2.8), then P satisfies $p^{u^+} = p^{u^-}$ for all $u \in \mathcal{N}$. By Lemma A.6 there are circuit vectors c_i such that $u = \sum_{i=1}^{l} c_i$, where the sum is sign-consistent. This means $u^+ = \sum_{i=1}^{r} c_i^+$ and $u^- = \sum_{i=1}^{l} c_i^-$. If P satisfies (2.8), then P satisfies $P^{c_i^+} - P^{c_i^-} = 0$. Using the equality

$$P^{u^+} - P^{u^-} = P^{\sum_{i=2}^{l} c_i^+} \left(P^{c_1^+} - P^{c_1^-} \right) + \left(P^{\sum_{i=2}^{l} c_i^+} - P^{\sum_{i=2}^{l} c_i^-} \right) P^{c_1^-},$$

the theorem follows by induction on l. □

Definition 2.26. Let $\mathcal{E}_1, \ldots, \mathcal{E}_c \subseteq \mathbf{P}(\mathcal{X})$. The *mixture* of $\mathcal{E}_1, \ldots, \mathcal{E}_c$ is the set of probability measures

$$\left\{ P = \sum_{i=1}^{c} \lambda_i P_i : P_1 \in \mathcal{E}_1, \ldots, P_c \in \mathcal{E}_c \text{ and } \lambda \in \mathbb{R}^c_{\geq}, \sum_{i=1}^{c} \lambda_i = 1 \right\}.$$

Corollary 2.27. *Let $\mathcal{E}$ be an exponential family with normal space $\mathcal{N}$. Let $\mathcal{Y} \subset \mathcal{X}$. If every circuit vector $c \in \mathcal{N}$ satisfies* $\operatorname{supp}(c) \subseteq \mathcal{Y}$ *or* $\operatorname{supp}(c) \subseteq \mathcal{X} \setminus \mathcal{Y}$*, then $\overline{\mathcal{E}}$ equals the mixture of $\overline{\mathcal{E}} \cap \mathbf{P}(\mathcal{Y})$ and $\overline{\mathcal{E}} \cap \mathbf{P}(\mathcal{X} \setminus \mathcal{Y})$.*

Proof. By Theorem 2.25, a probability measure $P \in \mathbf{P}(\mathcal{X})$ lies in $\overline{\mathcal{E}}$ if and only if its truncations $P^{\mathcal{Y}}$ and $P^{\mathcal{X}\setminus\mathcal{Y}}$ lie in $\overline{\mathcal{E}} \cap \mathbf{P}(\mathcal{Y})$ and $\overline{\mathcal{E}} \cap \mathbf{P}(\mathcal{X} \setminus \mathcal{Y})$, respectively. □

The corollary can be reformulated as follows, using terminology from matroid theory: If $\mathcal{X}_1, \ldots, \mathcal{X}_c$ are the connected components of the matroid of $\mathcal{N}$, then $\overline{\mathcal{E}}$ equals the mixture of $\overline{\mathcal{E}_1}, \ldots, \overline{\mathcal{E}_c}$, where $\mathcal{E}_i = \overline{\mathcal{E}} \cap \mathbf{P}(\mathcal{X}_i)^\circ$ is an exponential family on $\mathcal{X}_i$ for $i = 1, \ldots, c$.

The statement of Lemma 2.24 can be reversed:

Lemma 2.28. *Let $P_{\mathcal{E}}$ be the rI-projection of $P \in \mathbf{P}(\mathcal{X})$ onto an exponential family $\mathcal{E}$. Then* $\operatorname{supp}(P_{\mathcal{E}})$ *equals the smallest facial subset of $\mathcal{X}$ containing* $\operatorname{supp}(P)$*. In particular, a set $\mathcal{Y} \subseteq \mathcal{X}$ equals the support of some $Q \in \overline{\mathcal{E}}$ if and only if $\mathcal{Y}$ is facial.*

Proof. $\operatorname{supp}(P_{\mathcal{E}})$ contains $\operatorname{supp}(P)$, because otherwise $D(P\|P_{\mathcal{E}}) = \infty$. Since the intersection of faces of a polytope is again a face, the intersection of facial subsets is again facial. Hence there exists a smallest facial set $\mathcal{Y} \subseteq \mathcal{X}$ containing $\operatorname{supp}(P)$. Choose $\theta_0 \in \tilde{\mathcal{T}}$ and $a < b$ as in Definition 2.22. Let $x \in \mathcal{X} \setminus \mathcal{Y}$. Any $Q \in \mathbf{P}(\mathcal{X})$ with $Q(x) > 0$ satisfies $\sum_x \theta_0(x)Q(x) > a = \sum_x \theta_0(x)P(x)$, and hence $P - Q \notin \mathcal{N}$. Therefore, $\operatorname{supp}(P_{\mathcal{E}}) \subseteq \mathcal{Y}$. Since $\operatorname{supp}(P_{\mathcal{E}})$ is facial by Lemma 2.24 the assertion follows. □

Theorem 2.29. *Let $\mathcal{E}$ be an exponential family on $\mathcal{X}$ with reference measure ν and extended tangent space $\tilde{\mathcal{T}}$. Let $\mathcal{Y} \subset \mathcal{X}$ be facial, and let $\mathcal{E}^{\mathcal{Y}}$ be the set of all probability distributions P in $\overline{\mathcal{E}}$ such that* $\operatorname{supp}(P) = \mathcal{Y}$*. Let $\mathbf{1}_{\mathcal{Y}} : \mathcal{X} \to \{0, 1\}$ be the characteristic function of $\mathcal{Y}$. Then $\mathcal{E}^{\mathcal{Y}}$ consists of all probability measures of the form*

$$P_\theta^{\mathcal{Y}}(x) = \frac{\mathbf{1}_{\mathcal{Y}}(x)\nu_x}{Z_\vartheta^{\mathcal{Y}}} \exp(\theta(x)), \tag{2.9}$$

where $\theta \in \tilde{\mathcal{T}}$. In other words, $\mathcal{E}^{\mathcal{Y}}$ equals the set of truncations of $\mathcal{E}$ to $\mathcal{Y}$.

The proof makes use of the following lemma, which can also be generalized to a characterization of support sets; see [61].

Lemma 2.30. *Let $P \in \overline{\mathcal{E}}$ and $u \in \mathcal{N}$. Then* $\operatorname{supp}(u^+) \subseteq \operatorname{supp}(P)$ *if and only if* $\operatorname{supp}(u^-) \subseteq \operatorname{supp}(P)$.

Proof. Let $\mathcal{Y} = \operatorname{supp}(P)$. If $\operatorname{supp}(u) \not\subseteq \mathcal{Y}$, then take $x \in \operatorname{supp}(u) \setminus \mathcal{Y}$. Assume $u(x) > 0$. Plugging P and u into (2.6) the left hand side vanishes. Therefore the right hand side vanishes, too, and so there exists $y \in \operatorname{supp}(u) \setminus \mathcal{Y}$ such that $u(y) < 0$. □

Proof of Theorem 2.29. As above, assume $\nu = \mathbf{1}$ without loss of generality. By Theorem 2.21 a probability measure P with support $\mathcal{Y}$ belongs to $\overline{\mathcal{E}}$ if and only if P satisfies $P^{u^+} = P^{u^-}$ for all $u \in \mathcal{N}$. By Lemma 2.30, if $u \in \mathcal{N}$ satisfies $\operatorname{supp}(u) \not\subseteq \mathcal{Y}$, then $P^{u^+} = 0 = P^{u^-}$. It follows that a probability measure P with support $\mathcal{Y}$ belongs to $\overline{\mathcal{E}}$ if and only if P satisfies $P^{u^+} = P^{u^-}$ for all $u \in \mathcal{N}_\mathcal{Y} := \mathcal{N} \cap \mathbb{R}^\mathcal{Y}$. By Theorem 2.21 these equations characterize $\mathcal{E}^\mathcal{Y}$ as an exponential family over $\mathcal{Y}$ with normal space $\mathcal{N}_\mathcal{Y}$. Clearly $\tilde{\mathcal{T}}_\mathcal{Y} := \{\vartheta|_\mathcal{Y} : \vartheta \in \tilde{\mathcal{T}}\}$ is orthogonal to $\mathcal{N}_\mathcal{Y}$. Conversely, $\mathcal{N}_\mathcal{Y}^\perp = \mathcal{N}^\perp + \mathbb{R}^{\mathcal{X}\setminus\mathcal{Y}}$, so if $\theta' \in \mathbb{R}^\mathcal{Y}$ is orthogonal to $\mathcal{N}_\mathcal{Y}$, then $\theta' = \theta + \theta_2 = \theta|_\mathcal{Y}$, where $\theta \in \mathcal{N}^\perp = \tilde{\mathcal{T}}$ and $\theta_2 \in \mathbb{R}^{\mathcal{X}\setminus\mathcal{Y}}$. Therefore, $\mathcal{E}^\mathcal{Y}$ has the parametrization (2.9). □

2.3. Algebraic exponential families

For the basic definitions of commutative algebra and algebraic geometry the reader is referred to [19].

Definition 2.31. An exponential family $\mathcal{E}$ is called *algebraic* if its extended tangent space $\tilde{\mathcal{T}}$ is spanned (as a real vector space) by $\tilde{\mathcal{T}}_\mathbb{Z} := \tilde{\mathcal{T}} \cap \mathbb{Z}^\mathcal{X}$. In other words, $\mathcal{E}$ is algebraic if and only if it has an integer valued sufficient statistics matrix $A \in \mathbb{Z}^{h\times\mathcal{X}}$. Equivalently, $\mathcal{E}$ is algebraic if and only if its normal space $\mathcal{N}$ is spanned by $\mathcal{N}_\mathbb{Z} := \mathcal{N} \cap \mathbb{Z}^\mathcal{X}$.

The equivalence follows from elementary facts about homogeneous systems of linear equations with integer coefficients: Namely, the solution space of such a system has a basis that consists of integral vectors. As a consequence of this, algebraic exponential families can be described as the intersection of an algebraic subvariety of $\mathbb{C}^\mathcal{X}$ with $\mathbf{P}(\mathcal{X})^\circ$:

Theorem 2.32 (Geiger, Meek, Sturmfels [31])**.** *Let $\mathcal{E}$ be an algebraic exponential family with normal space $\mathcal{N}$. A probability distribution $P \in \mathbf{P}(\mathcal{X})$ belongs to $\overline{\mathcal{E}}$ if and only if P satisfies the polynomial equations*

$$P^{u^+}\nu^{u^-} = P^{u^-}\nu^{u^+}, \qquad \text{for all } u \in \mathcal{N}_\mathbb{Z}. \tag{2.10}$$

Proof. This theorem is an easy consequence of Theorem 2.25 and the following fact: If the vector space $\mathcal{N}$ has an integral basis, then $\mathcal{N}$ has a circuit basis that consists of integer vectors. The reason is that in this case the set of circuit vectors belonging to a given circuit $\mathcal{Y} \subseteq \mathcal{X}$ can be characterized by linear equations with integer coefficients. □

This connection between exponential families and algebraic geometry is one of the most fruitful topics in the relatively new field of algebraic statistics, which tries to systematically analyze statistical problems with the help of algebraic methods, see [57] and [26]. Algebraic exponential families are particularly nice, because many questions about algebraic exponential families can be answered by solving algebraic equations. This means that the tools of commutative algebra and algebraic geometry are available to study algebraic exponential families. For example, the rI-projection $P_{\mathcal{E}}$ of some $P \in \mathbf{P}(\mathcal{X})$ is characterized by the algebraic conditions $P_{\mathcal{E}} \in \overline{\mathcal{E}}$ and $P_{\mathcal{E}} \in \mathcal{N}_P$. Unfortunately, it is still not possible to find an analytical expression for the map $P \mapsto P_{\mathcal{E}}$: For example, Proposition 3 in [31] shows that, in general, the rI-projection has no closed form in terms of radicals.

There are a lot of specialized algorithms and computer algebra systems for commutative algebra, like Singular [34] and Macaulay2 [32]. Sections 3.6 and 3.7 contain two examples that show the power of computer algebra systems.

In the following assume that $\mathcal{N}$ is spanned by $\mathcal{N}_{\mathbb{Z}}$. Polynomial equations are easier to understand over an algebraically closed field. For this reason, equations (2.10) will be considered as complex polynomials in the following. Denote by $\mathbb{C}[P(x) : x \in \mathcal{X}]$ the polynomial ring with one variable $P(x)$ for each $x \in \mathcal{X}$. For any subset $B \subseteq \mathcal{N}_{\mathbb{Z}}$ let $I_\nu(B)$ be the ideal in $\mathbb{C}[P(x) : x \in \mathcal{X}]$ generated by the polynomials

$$P^{u^+}\nu^{u^-} - P^{u^-}\nu^{u^+}, \qquad \text{for all } u \in B.$$

For all algebraic considerations the reference measure plays only a secondary role: There is a linear coordinate transformation $\phi : u(x) \mapsto \frac{u(x)}{\nu_x}$ such that $\phi^{-1}(I_\nu(B)) = I_{\mathbf{1}}(B)$ (this fact was used already several times, e.g. in the proof of Theorem 2.21).

Remark 2.33. Elements from an exponential family also satisfy $\sum_x P(x) = 1$. This equation can be seen as a normalization condition. The ideal $I_\nu(B)$ is generated by homogeneous polynomial equations. This means that any positive solution $\mu \in \mathbb{R}^{\mathcal{X}}_{>}$ to $I_\nu(B)$ induces a normalized positive solution $\frac{1}{\mu(\mathcal{X})}\mu \in \mathbf{P}(\mathcal{X})$. Formally, the ideal $I_\nu(B)$ defines a projective variety, and the condition $\sum_x P(x) \neq 0$ defines an affine subvariety that contains the closure of the exponential family. For some theoretical considerations it is better to work in this projective picture, but in applications it is convenient to use the normalization condition $\sum_x P(x) = 1$ to eliminate one of the variables $P(x)$.

The ideals $I_\nu(B)$ are binomial ideals, i.e. they are defined by polynomials that have only two terms. Binomial ideals are special in many ways, and there are specialized algorithms to deal with them, due to the fact that a lot of algebraic properties of binomial ideals can be interpreted combinatorially. There is a specialized Macaulay2

package for computations with binomial ideals [43]. The theory of binomial ideals started with the paper [27] of Eisenbud and Sturmfels. One of the main results is the characterization of binomial prime ideals. The following construction is needed:

Definition 2.34. Let $\mathcal{X}$ be a finite set. A subgroup of $\mathbb{Z}^{\mathcal{X}}$ is called a *lattice*. A lattice $\mathcal{L}$ is *saturated* if there exists a vector space $\mathcal{N} \subseteq \mathbb{R}^{\mathcal{X}}$ such that $\mathcal{L} = \mathcal{N}_{\mathbb{Z}} := \mathcal{N} \cap \mathbb{Z}^{\mathcal{X}}$. A *character* of a lattice $\mathcal{L}$ is a group homomorphism from $\mathcal{L}$ into the multiplicative group $\mathbb{C}^{\times}$ of $\mathbb{C}$.

Let $\mathcal{Y} \subseteq \mathcal{X}$, let $\mathcal{L} \subseteq \mathbb{Z}^{\mathcal{Y}}$ be a lattice, and let $\rho : \mathcal{L} \to \mathbb{C}^{\times}$ be a character. For any subset $B \subseteq \mathcal{L}$ let $I_{\mathcal{Y},\mathcal{L},\rho}(B) \subset \mathbb{C}[P(x) : x \in \mathcal{X}]$ be the ideal generated by

$$\begin{aligned} P(x), &\qquad \text{for all } x \notin \mathcal{Y}, \\ P^{v+} - \rho(v)P^{v-}, &\qquad \text{for all } v \in B, \end{aligned}$$

Denote $I_{\mathcal{Y},\mathcal{L},\rho}(\mathcal{L})$ by $I_{\mathcal{Y},\mathcal{L},\rho}$. Let $\rho_1 : v \mapsto 1$ be the trivial character. The ideal $I_{\mathcal{Y},\mathcal{L},\rho_1}$ is also called the *lattice ideal* of $\mathcal{L}$. If ρ is an arbitrary character, then the linear change of coordinates ϕ defined above transforms $I_{\mathcal{Y},\mathcal{L},\rho}(C)$ into a lattice ideal.

Theorem 2.35 (Eisenbud, Sturmfels [27])**.** *A binomial ideal $I \subseteq \mathbb{C}[P(x) : x \in \mathcal{X}]$ is prime if and only if it is of the form $I = I_{\mathcal{Y},\mathcal{L},\rho}$, where $\mathcal{L} \subseteq \mathbb{Z}^{\mathcal{Y}}$ is a saturated lattice.*

Proof. See [27, Corollary 2.6]. □

It follows from Theorem 2.35 that $I_\nu(\mathcal{N}_{\mathbb{Z}})$ is a prime ideal: In this case $\mathcal{Y} = \mathcal{X}$, $\mathcal{L} = \mathcal{N}_{\mathbb{Z}}$, and ρ is given by $\rho(v) = \nu^v$. Therefore, the corresponding variety $V_{\mathcal{E}} := V(I_\nu(\mathcal{N}_{\mathbb{Z}}))$ is irreducible, i.e. it cannot be written as a union $V_1 \cup V_2$ of nontrivial subvarieties. Since the (real) dimension of $\mathcal{E}$ equals the complex dimension of $V_{\mathcal{E}}$, it follows that $V_{\mathcal{E}}$ is the Zariski closure of $\mathcal{E}$, i.e. the smallest variety containing $\overline{\mathcal{E}}$.

Definition 2.36. A binomial prime ideal is also called a *toric ideal*. The (affine or projective) variety of a toric ideal is called an (affine or projective) *toric variety*.

Toric varieties are an important subject of algebraic geometry. See [30] and [18] for an introduction.

Definition 2.37. Let $\mathcal{Y}$, $\mathcal{L}$ and ρ be as before. A finite set $B \subset \mathcal{L}$ is called a *Markov basis* of $I_{\mathcal{Y},\mathcal{L},\rho}$ if $I_{\mathcal{Y},\mathcal{L},\rho} = I_{\mathcal{Y},\mathcal{L},\rho}(B)$.

Hilbert's Basissatz implies the existence of Markov bases. A change of coordinates similar to ϕ defined above can be used to prove that the notion of a Markov basis does not depend on ρ. The name "Markov bases" comes from the fact that a Markov basis can be used as a set of moves in a Markov Chain Monte Carlo simulation to explore the integer points of polytopes, see [25].

It follows from the proof of Theorem 2.32 that it would be enough to consider the polynomial equations coming from an integral circuit basis. Therefore, a circuit basis is a natural candidate for a Markov basis. Yet a Markov basis may contain vectors that

are not circuit vectors. An example is given by the binomial models $\mathrm{Bin}(n)$ for $n \geq 3$, see [61]. Even in the case that a given integral circuit basis C is not a Markov basis, Theorem 2.25 implies that the ideals $I_\nu(C)$ and $I_\nu(\mathcal{N}_\mathbb{Z})$ have the same non-negative real solutions. Still, there are some computational issues to bear in mind.

Over $\mathbb{C}$ the equations corresponding to different integral circuit bases do not have the same solutions. The reason is that proportional circuit vectors only yield equivalent equations if they are considered over the non-negative real numbers. In particular, in general $I_\nu(C) \neq I_\nu(C')$ if C and C' are two different circuit bases. From a computational viewpoint it is advantageous to choose the equations such that the number of negative real solutions or complex solutions is as small as possible. This can be achieved by considering those integral circuit vectors that are as small as possible:

Definition 2.38. An integer vector $u \in \mathbb{Z}^\mathcal{X}$ is *prime* if the greatest common divisor $\gcd\{u(x) : x \in \mathcal{X}\}$ of its components is one. A circuit basis is *prime* if all of its elements are integral and prime.

The prime circuit vector corresponding to a given circuit is uniquely determined up to a sign. Therefore, in the algebraic case the distinction between circuits and circuit vectors is often ignored.

Any integral circuit vector is an integral multiple of a prime circuit vector, so prime circuit bases exist. Furthermore, if C is a prime circuit basis and $\overline{C}$ is the set of all integral circuits, then $I_\nu(\overline{C}) = I_\nu(C)$. If C is a prime circuit basis, then $I_\mathbf{1}(C)$ is also called the *circuit ideal* of $\mathcal{N}$. If ν is arbitrary, then the linear change defined above of coordinates ϕ transforms $I_\nu(C)$ into a circuit ideal.

The question when a circuit ideal is a toric ideal is discussed in [14]. In general, the difference between toric ideals and circuit ideals is, in a certain sense, small:

Proposition 2.39. *Let $\mathcal{Y} \subseteq \mathcal{X}$, let B be a Markov basis of a saturated lattice $\mathcal{L} \subseteq \mathbb{Z}^\mathcal{Y}$, and let $\overline{C}$ be the set of circuits of $\mathcal{L}$. For any character ρ on $\mathcal{L}$ the ideal $I_{\mathcal{Y},\mathcal{L},\rho}(B)$ equals the radical of $I_{\mathcal{Y},\mathcal{L},\rho}(\overline{C})$.*

Proof. See Proposition 8.7 in [27]. □

Corollary 2.40. *If $\mathcal{E}$ is an algebraic exponential family and if C is a prime circuit basis of its normal space $\mathcal{N}$, then $I_\nu(\mathcal{N}_\mathbb{Z})$ equals the radical of $I_\nu(C)$. In particular, the variety of $I_\nu(C)$ equals the Zariski closure of $\mathcal{E}$.*

Corollary 2.40 shows that the ideal $I_\nu(C)$ of any prime circuit basis C does not admit any superfluous non-negative or complex solutions. Still, knowing the radical of an ideal may greatly decrease the running time of many algorithms of computational commutative algebra, so it is preferable to work with a Markov basis instead of a prime circuit basis if possible.

Finding a Markov basis or a circuit basis is in general a non-trivial task. [48] and [36] discuss algorithms for both tasks, which are implemented in the open source software package 4ti2 [1]. Let A be a matrix such that $\mathcal{N}$ equals the row space of A. Markov basis computations tend to depend on the size of the entries of A: If A has only small

entries, then one may hope that there are "enough" vectors in $\mathcal{N}_\mathbb{Z}$ with small entries, corresponding to polynomials of low degree. The Markov bases algorithm is related to Buchberger's algorithm, and the speed of this algorithm depends on the degrees of the starting polynomials. Circuit computations do not depend in any essential way on the size of the entries of A. The number of circuits, however, tends to be much larger than the number of Markov basis elements. Therefore, generically, Markov basis computations are faster when A has only "small" entries (which is the most important case for applications), and circuit computations are faster when A has "large" entries.

Most algorithms to compute Markov bases make use of a relation similar to the relation stated in the following proposition. First, a definition is needed:

Definition 2.41. Let $I \subseteq \mathbb{C}[P(x) : x \in \mathcal{X}]$ be an ideal and let $J \subseteq \mathbb{C}[P(x) : x \in \mathcal{X}]$ be an arbitrary set of polynomials. Then the ideal

$$I : J = \{f \in \mathbb{C}[P(x) : x \in \mathcal{X}] : fg \in I \text{ for all } g \in J\}$$

is called the *quotient* of I with respect to J. The ideal

$$I : J^\infty := \bigcup_{n \geq 1} I : J^n,$$

where $I : J^1 = I : J$ and $I : J^{n+1} = (I : J^n) : J$, is called the *saturation* of I with respect to J.

The saturation of two ideals I, J corresponds to the difference of the varieties, in the sense that the variety $V(I : J)$ of $I : J$ equals the Zariski closure of $V(I) \setminus V(J)$.

Proposition 2.42. *Let $\mathcal{Y} \subseteq \mathcal{X}$, let $\mathcal{L} \subseteq \mathbb{Z}^\mathcal{Y}$ be a saturated lattice, and let ρ be a character of $\mathcal{L}$. If $B \subseteq \mathcal{L}$ generates $\mathcal{L}$ (as an abelian group), then*

$$I_{\mathcal{Y},\mathcal{L},\rho} = I_{\mathcal{Y},\mathcal{L},\rho}(B) : (\prod_{x \in \mathcal{Y}} p(x))^\infty.$$

Hence, the varieties $V_{\mathcal{Y},\mathcal{L},\rho}$ and $V_{\mathcal{Y},\mathcal{L},\rho}(B)$ of $I_{\mathcal{Y},\mathcal{L},\rho}$ and $I_{\mathcal{Y},\mathcal{L},\rho}(B)$ satisfy

$$V_{\mathcal{Y},\mathcal{L},\rho} \cap \left(\mathbb{C}^{\mathcal{X}\setminus\mathcal{Y}} \oplus (\mathbb{C}^\times)^\mathcal{Y}\right) = V_{\mathcal{Y},\mathcal{L},\rho}(B) \cap \left(\mathbb{C}^{\mathcal{X}\setminus\mathcal{Y}} \oplus (\mathbb{C}^\times)^\mathcal{Y}\right),$$

where $\mathbb{C}^\times = \mathbb{C} \setminus \{0\}$.

Proof. See [69, Lemma 12.2]. □

The second statement of the proposition, applied to algebraic exponential families, is the algebraic version of the fact that a basis of $\mathcal{N}$ is enough to describe $\mathcal{E}$: A probability measure $P \in \mathbf{P}(\mathcal{X})^\circ$ lies in $\mathcal{E}$ if and only if $\log(P)$ is orthogonal to (a basis of) $\mathcal{N}$ (cf. the proof of Theorem 2.21). Things get complicated only at the boundary.

2.4. Hierarchical models

Hierarchical models are important examples of algebraic exponential families, and many examples appearing throughout this thesis are, in fact, hierarchical models. They describe the interaction of a finite set of finite subsystems $\mathcal{X}_i$. The restrictions $X_i : \mathcal{X} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathcal{X}_i$ to the subsystems can be viewed as random variables, and hierarchical models can be used to study the relationship of these discrete random variables. This section summarizes the main facts which are needed in the following. See [46] and [26] for further information.

Unfortunately, different authors mean different things when they talk about hierarchical models. Here, the following definition is used:

Definition 2.43. Let $\mathcal{X}_1, \ldots, \mathcal{X}_n$ be finite sets of cardinality $|\mathcal{X}_i| = N_i$, and let $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$. For any subset $S \subseteq [n]$ let $\mathcal{X}_S = \times_{i \in S} \mathcal{X}_i$. For any family Δ of subsets of $[n]$ let $\mathcal{E}'_\Delta$ be the set of all probability measures $P \in \mathbf{P}(\mathcal{X})^\circ$ that can be written in the form

$$P(x) = \prod_{S \in \Delta} f_S(x), \tag{2.11}$$

where each f_S is a non-negative function on $\mathcal{X}$ that depends only on those components of x lying in S. In other words, $f_S(x) = f_S(y)$ for all $x = (x_i)_{i=1}^n, y = (y_i)_{i=1}^n \in \mathcal{X}$ satisfying $x_i = y_i$ for all $i \in S$. The *hierarchical exponential family* $\mathcal{E}_\Delta$ of Δ with parameters N_1, N_2, …, N_n is defined as $\mathcal{E}'_\Delta \cap \mathbf{P}(\mathcal{X})^\circ$. The closure of $\mathcal{E}_\Delta$ (which equals the closure of $\mathcal{E}'_\Delta$) is called the *hierarchical model* of Δ with parameters N_1, N_2, …, N_n. If $N_i = N_1$ for all $i = 1, \ldots, n$, then the hierarchical model and the hierarchical exponential family are called *homogeneous of size* N_1. They are called *binary* if they are homogeneous of size two.

At first sight one might think that $\mathcal{E}'_\Delta = \overline{\mathcal{E}_\Delta}$. Unfortunately, this is not true, see [31]. For certain applications, when the factorizability probability is important, one might want to call $\mathcal{E}'_\Delta$ a hierarchical model. When studying optimization problems it is more important that the models are closed.

If $S \in \Delta$ and if $S' \subset S$, then Δ and $\Delta \cup \{S'\}$ determine the same hierarchical model. In order to make the correspondence $\Delta \mapsto \mathcal{E}_\Delta$ injective, it is convenient to require that Δ is a *simplicial complex*. By definition this means that $S \in \Delta$ and $S' \subseteq S$ implies $S' \in \Delta$. This requirement associates the largest possible set Δ to a hierarchical model $\mathcal{E}$. Alternatively, there is a unique minimal set Δ describing $\mathcal{E}$; it is characterized by the following property: If $S, S' \in \Delta$ satisfy $S' \subseteq S$, then $S' = S$. Due to its minimality property this second choice is convenient when giving examples.

For any $S \subseteq \{1, \ldots, n\}$ the subset of $\mathbb{R}^\mathcal{X}$ of functions that only depend on the S-components can be naturally identified with $\mathbb{R}^{\mathcal{X}_S}$. The natural projection $\mathcal{X} \to \mathcal{X}_S$ induces a natural injection $\mathbb{R}^{\mathcal{X}_S} \to \mathbb{R}^\mathcal{X}$.

It is easy to see that hierarchical exponential families are indeed exponential families: Namely, (2.11) implies that $\mathcal{E}_\Delta$ consists of all $P \in \mathbf{P}(\mathcal{X})^\circ$ that satisfy

$$(\log(P(x)))_{x \in \mathcal{X}} \in \sum_{S \in \Delta} \mathbb{R}^{\mathcal{X}_S} \subseteq \mathbb{R}^\mathcal{X}.$$

Therefore, $\mathcal{E}_\Delta$ is an exponential family with uniform reference measure and extended tangent space $\tilde{\mathcal{T}} = \sum_{S\in\Delta} \mathbb{R}^{\mathcal{X}_S}$. This vector space sum is not direct, since every summand contains $\mathbf{1}$. There is a natural sufficient statistics: The marginalization maps $\pi_S : \mathbb{R}^{\mathcal{X}} \mapsto \mathbb{R}^{\mathcal{X}_S}$ defined for $S \subseteq \{1, \ldots, n\}$ via

$$\pi_S(v)(x) = \sum_{y\in\mathcal{X}: y_i = x_i \text{ for all } i\in S} v(y)$$

induce the moment map

$$\pi_\Delta : v \in \mathbb{R}^{\mathcal{X}} \mapsto (\pi_S(v))_{S\in\Delta} \in \bigoplus_{S\in\Delta} \mathbb{R}^{\mathcal{X}_S},$$

where $\oplus$ denotes the external direct sum of vector spaces.

Definition 2.44. For any $S \subseteq [n]$ the image $\pi_S(v)$ of $v \in \mathbb{R}^{\mathcal{X}}$ under π_S is called the *S-marginal* of v. If Δ is a collection of subsets of $[n]$ and π_Δ is defined as above, then $\pi_\Delta(v) = (\pi_S(v))_{S\in\Delta}$ is called the *Δ-marginal* of v. The convex support of the hierarchical model $\overline{\mathcal{E}_\Delta}$ is called the *marginal polytope* (of $\overline{\mathcal{E}_\Delta}$ or of Δ).

Lemma 2.45. *Let Δ be a collection of subsets of $[n]$, and let $K = \cup_{J\in\Delta} J$. The marginal polytope of Δ is (affinely equivalent to) a 0-1-polytope with $\prod_{i\in K} N_i$ vertices.*

Proof. The moment map π_Δ corresponds to a sufficient statistics A_Δ that only has entries 0 and 1, so $\mathbf{M}_A$ is a 0-1-polytope. The set of vertices of $\mathbf{M}_A$ is a subset of $\{A_x : x \in \mathcal{X}\}$. Let $x = (x_i)_{i=1}^n, y = (y_i)_{i=1}^n \in \mathcal{X}$. If $x_i = y_i$ for all $i \in K$, then $A_x = A_y$, so $\mathbf{M}_A$ has at most $\prod_{i\in K} N_i$ vertices. If $x_i \neq y_i$ for some $i \in K$, then $A_x \neq A_y$, so the set $\{A_x : x \in \mathcal{X}\}$ has cardinality $\prod_{i\in K} N_i$. Since this set consists of 0-1-vectors and since no 0-1-vector is a convex combination of other 0-1-vectors, it follows that the set of vertices of $\mathbf{M}_A$ equals $\{A_x : x \in \mathcal{X}\}$ and has cardinality $\prod_{i\in K} N_i$. □

The following examples of hierarchical models are particularly important:

Definition 2.46. For $1 \leq k \leq n$ let $\Delta_k = \{J \subseteq [n] : |J| = k\}$. The hierarchical model of Δ_k is called the *k-interaction model*. The hierarchical models of Δ_1, Δ_2 and Δ_3 are also called (in a slightly inconsistent manner) the *independence model*, the *pair interaction model* and the *three-way interaction model*, respectively.

Denote by X_i the random variable corresponding to the natural projection $\mathcal{X} \mapsto \mathcal{X}_i$. If P is the joint distribution of $X_1, \ldots, X_n$, then the information divergence $D(P\|\mathcal{E}_1)$ is called the *multiinformation* of $X_1, \ldots, X_n$, denoted by $I(X_1; \ldots; X_n)$. If $n = 2$, then $I(X_1; X_2)$ is also called the *mutual information* of X_1 and X_2.

A basis of the normal space $\mathcal{N}$ can be found using the following construction from [38]. This basis has the advantage that it has only few nonzero entries, and these are all equal to plus or minus one.

Definition 2.47. Assume that $\mathcal{X}_i = \{1, \dots, N_i\}$. Let S be a subset of $[n]$. For each $j \in S$ let $1 \leq i_j < N_j$. Then the vector $u_{S;\{i_j\}_{j\in S}} \in \mathbb{R}^{\mathcal{X}}$ defined by

$$u_{S;\{i_j\}_{j\in S}}(x_1, \dots, x_n) = \begin{cases} (-1)^{\sum_{j\in S}(x_j - i_j)}, & \text{if } x_j \in \{i_j, i_j + 1\} \text{ for all } j \in S \\ & \qquad \text{and } x_j = 1 \text{ else,} \\ 0, & \text{otherwise.} \end{cases}$$

is called an *adjacent minor supported on* S.

Theorem 2.48. *Let Δ be a simplicial complex. The set of all adjacent minors which are supported on some $S \notin \Delta$ forms a basis of the normal space $\mathcal{N}$ of the hierarchical model $\mathcal{E}_\Delta$. Therefore, the dimension of $\mathcal{E}_\Delta$ is*

$$\sum_{S\in\Delta} \prod_{i\in S} (N_i - 1) - 1.$$

Proof. See Theorem 2.6 and Corollary 2.7 in [38]. □

Examples 3.13 and 3.42 contain the bases of adjacent minors for the independence model of two binary variables and the pair interaction model of four binary variables.

Let $S \subseteq [n]$. It is sometimes useful to separate the "pure" S-interactions from the lower S'-interactions for S' contained in S. This amounts to choosing an orthogonal complement to $\sum_{S'\subsetneq S} \mathbb{R}^{\mathcal{X}_{S'}}$ in $\mathbb{R}^{\mathcal{X}_S}$. A natural choice is the orthogonal complement (with respect to the natural scalar product). For example, the *exponential family of pure pair interactions* $\mathcal{E}_{(2)}$ has the uniform reference measure and the tangent space

$$\mathcal{T}_{(2)} = \left\{ \theta \in \sum_{S\subseteq[n],|S|=2} \mathbb{R}^{\mathcal{X}_S} : \sum_{x\in\mathcal{X}} \theta(x) f(x) = 0 \text{ for all } i \in [n] \text{ and all } f \in \mathbb{R}^{\mathcal{X}_i} \right\} + \mathbb{R}\mathbf{1}.$$

See Section 5.1 for another possibility to quantify the contributions of different levels of interaction, which uses the Pythagorean identity.

3. Maximizing the information divergence from an exponential family

Nihat Ay proposed the following mathematical problem in [5]:

- Given an exponential family $\mathcal{E}$, find the maximal value of $D_{\mathcal{E}} := D(\cdot\|\mathcal{E})$, and find the maximizing probability measures.

An overview of previous works on this problem as well as a short summary of the applications was already given in the introduction. Some of the previous results will be summarized in Section 3.1. The possible applications will be discussed again in Chapter 5, after the problem has been studied from its mathematical side.

The main result in this chapter, Theorem 3.28, shows that the original maximization problem can be solved by studying the related problem:

- Maximize the function $\overline{D}_{\mathcal{E}}(u) = \sum_{x\in X} u(x) \log \frac{|u(x)|}{\nu_x}$ for all $u \in \mathcal{N}$ such that $\sum_x |u(x)| \leq 2$.

The local and global maximizers of $\overline{D}_{\mathcal{E}}$ are in bijection with the local and global maximizers of $D_{\mathcal{E}}$.

Section 3.1 presents Matúš's result on the directional derivatives of $D_{\mathcal{E}}$ and some corollaries, including the first order optimality conditions of maximizers of $D_{\mathcal{E}}$. These conditions include the projection property, which is analyzed in Section 3.2, leading to the notion of a kernel distribution. It is easy to see that probability measures that satisfy the projection property and that do not belong to $\mathcal{E}$ come in pairs (P^+, P^-) such that $P^+ - P^- \in \mathcal{N}$ and $\operatorname{supp}(P^+) \cap \operatorname{supp}(P^-) = \emptyset$. This pairing is used in Section 3.3 to relate the maximization of $D_{\mathcal{E}}$ to the maximization of $\overline{D}_{\mathcal{E}}$. The first order optimality conditions of $\overline{D}_{\mathcal{E}}$ are computed in Section 3.4. Section 3.5 contains the main result of this chapter, Theorem 3.28, which states that finding the critical points or the local or global maximizers of $D_{\mathcal{E}}$ is equivalent to finding the critical points or the local or global maximizers of $\overline{D}_{\mathcal{E}}$. Section 3.6 discusses how to solve the critical equations from Section 3.4. Section 3.7 presents an alternative method for computing the local maximizers of $D_{\mathcal{E}}$, which uses the projection property more directly. Sections 3.6 and 3.7 contain two examples where the global maximizers were unknown before. Some of the results in this section have been published in [60] and [53].

3.1. The directional derivatives of $D(\cdot\|\mathcal{E})$

The directional derivatives of the information divergence where first computed by Ay in the special case that $P_{\mathcal{E}}$ has full support, see [5]. The general case is due to Matúš:

Theorem 3.1. *Let $\mathcal{E}$ be an exponential family on a finite set $\mathcal{X}$, and let $P, R \in \mathbf{P}(\mathcal{X})$. Let $\mathcal{Y} \subseteq \mathcal{X}$ be the smallest facial set containing* $\operatorname{supp}(P)$.

- *If* $\operatorname{supp}(R) \subseteq \operatorname{supp}(P)$, *then the two-sided derivative of $D_{\mathcal{E}}$ at P in the direction $R - P$ equals*
$$\sum_{x \in \operatorname{supp}(P)} [R(x) - P(x)] \log \frac{P(x)}{P_{\mathcal{E}}(x)}.$$
- *If* $\operatorname{supp}(R) \subseteq \mathcal{Y}$ *and if $R(\mathcal{X} \setminus \operatorname{supp}(P)) > 0$, then the one-sided derivative of $D_{\mathcal{E}}$ at P in the direction $R - P$ equals $-\infty$.*

Proof. See Theorem 4.1 in [50]. □

Theorem 4.1 in [50] also discusses the one-sided directional derivatives in the case $\operatorname{supp}(R) \setminus \mathcal{Y} \neq \emptyset$. From this more general result the following theorem can be deduced:

Theorem 3.2. *Let P be a local maximizer of $D_{\mathcal{E}}$ with support $\mathcal{Z} = \operatorname{supp}(P)$, and let $P_{\mathcal{E}}$ be its rI-projection onto $\mathcal{E}$. Then the following holds:*

(*i*) *Suppose $\mathcal{Y} := \operatorname{supp}(P_{\mathcal{E}}) \neq \mathcal{X}$. Then $\{A_x : x \in \mathcal{Y}\}$ and $\{A_x : x \in \mathcal{X} \setminus \mathcal{Y}\}$ lie in distinct parallel hyperplanes.*

(*ii*) *P satisfies the* projection property, *i.e. P equals the truncation $P_{\mathcal{E}}^{\mathcal{Z}}$ of $P_{\mathcal{E}}$ to $\mathcal{Z}$:*
$$P(x) = \begin{cases} \frac{P_{\mathcal{E}}(x)}{P_{\mathcal{E}}(\mathcal{Z})}, & \text{if } x \in \mathcal{Z}, \\ 0, & \text{else.} \end{cases} \tag{3.1}$$

(*iii*) *Assume $\mathcal{Y} \neq \mathcal{X}$, and let $\mathcal{E}^{P_{\mathcal{E}}} = \{Q^{\mathcal{X}\setminus\mathcal{Y}} : Q \in \mathcal{E} \text{ and } Q^{\mathcal{Y}} = P_{\mathcal{E}}\}$. Then*
$$D(P\|\mathcal{E}) \geq D(R\|\mathcal{E}^{P_{\mathcal{E}}}) \qquad \text{for all probability measures } R \text{ on } \mathcal{X} \setminus \mathcal{Y}. \tag{3.2}$$

Proof. Statement (*ii*) is due to Ay [5] in the special case where $\mathcal{Y} = \mathcal{X}$. Statements (*i*) and (*iii*) and the general form of statement (*ii*) are due to Matúš [50, Theorem 5.1]. □

Definition 3.3. A probability measure that satisfies the projection property (3.1) is called a *projection point*. A projection point P is *proper* if $P \notin \overline{\mathcal{E}}$. A proper projection point that satisfies statement (*i*) of Theorem 3.2 is called a *quasi-critical point* of $D_{\mathcal{E}}$. If $P \in \mathbf{P}(\mathcal{X}) \setminus \overline{\mathcal{E}}$ satisfies all three conclusions of Theorem 3.2, then P is a *critical point* of $D_{\mathcal{E}}$.

This definition is motivated by the following philosophy: In convex analysis, a point satisfying all first-order optimality conditions (which in general comprise both equations and inequalities) of an optimization problem is called a *critical point*. In many respects, equations are much easier to analyze than inequalities. Therefore it makes sense to treat these two kinds of conditions separately, and the term "quasi-critical" point is chosen for points that satisfy only the equations derived from the first order optimality conditions. It is convenient to exclude the trivial solutions $P \in \overline{\mathcal{E}}$ in these definitions.

Lemma 3.4. *Let $\mathcal{E}$ be an exponential family with normal space $\mathcal{N}$, and let $Q \in \overline{\mathcal{E}} \setminus \mathcal{E}$. Write $\mathcal{Y} = \operatorname{supp}(Q)$ and $\mathcal{Y}' = \mathcal{X} \setminus \mathcal{Y}$. Then $\mathcal{E}^Q = \{P^{\mathcal{Y}'} : P \in \mathcal{E} \text{ and } P^{\mathcal{Y}} = Q\}$ is an exponential family on $\mathcal{Y}'$ with normal space $\mathcal{N}^{\mathcal{Y}'} = \{v|_{\mathcal{Y}'} : v \in \mathcal{N}, \sum_{x\in\mathcal{Y}'} v(x) = 0\}$.*

Proof. The statements of this lemma are hidden in the proof of Theorem 5.1 in [50]. Let $\tilde{\mathcal{T}}$ be the extended tangent space of $\mathcal{E}$, and let $\mathcal{E}' = \{P \in \mathcal{E} : P^{\mathcal{Y}} = Q\}$. Then $\mathcal{E}'$ is an exponential family with extended tangent space

$$\tilde{\mathcal{T}}' = \left\{\theta \in \tilde{\mathcal{T}} : \theta|_{\mathcal{Y}} \in \mathbb{R}\mathbf{1}_{\mathcal{Y}}\right\} = \left(\tilde{\mathcal{T}} \cap \mathbb{R}^{\mathcal{Y}'}\right) + \mathbb{R}\mathbf{1}.$$

Therefore, $\mathcal{E}^Q = \{P^{\mathcal{Y}'} : P \in \mathcal{E}'\}$ is also an exponential family, with normal space

$$\mathcal{N}^{\mathcal{Y}'} = \left\{v|_{\mathcal{Y}'} : v \in \mathbb{R}^{\mathcal{X}} \text{ is orthogonal to } \tilde{\mathcal{T}}', \sum_{x\in\mathcal{Y}'} v(x) = 0\right\},$$

hence $\mathcal{N}^{\mathcal{Y}'} \supseteq \{v|_{\mathcal{Y}'} : v \in \mathcal{N}, \sum_{x\in\mathcal{Y}'} v(x) = 0\}$. The orthogonal complement of $\tilde{\mathcal{T}}'$ is $\left(\tilde{\mathcal{T}}^\perp + \left(\mathbb{R}^{\mathcal{Y}'}\right)^\perp\right) \cap \mathbf{1}_{\mathcal{Y}}^\perp \subseteq \mathcal{N} + \mathbb{R}^{\mathcal{Y}}$. Therefore, any $w \in \mathcal{N}^{\mathcal{Y}'}$ is of the form $(v_1 + v_2)|_{\mathcal{Y}'} = v_1|_{\mathcal{Y}'}$, where $v_1 \in \mathcal{N}$ and $v_2 \in \mathbb{R}^{\mathcal{Y}}$, so $\mathcal{N}^{\mathcal{Y}'} \subseteq \{v|_{\mathcal{Y}'} : v \in \mathcal{N}, \sum_{x\in\mathcal{Y}'} v(x) = 0\}$. □

The following two lemmas are useful to restrict the study of the maximizers of $D_{\mathcal{E}}$ to small faces of the probability simplex $\mathbf{P}(\mathcal{X})$. The first lemma is due to Ay [5] and Matúš [52, Proposition 3.2]. It is instructive to compare its proof to the proof of Lemma 3.30.

Lemma 3.5. *Let $\mathcal{E}$ be an exponential family with normal space $\mathcal{N}$ and sufficient statistics A. Let P be a local maximizer of $D_{\mathcal{E}}$, and let $\mathcal{Z} = \operatorname{supp}(P)$. Then the set $\{A_x : x \in \mathcal{Z}\}$ is affinely independent. Equivalently, if $v \in \mathcal{N}$ satisfies $\operatorname{supp}(v) \subseteq \mathcal{Z}$, then $v = 0$. In particular, the cardinality of $\mathcal{Z}$ is bounded by $\dim \mathcal{E} + 1$.*

Proof. By Theorem 2.16 the restriction of $D_{\mathcal{E}}$ to the polytope $\mathcal{N}_P$ is strictly convex. If P is a local maximizer of $D_{\mathcal{E}}$, then P must be an extreme point of $\mathcal{N}_P$. Assume that $v \in \mathcal{N}$ satisfies $\operatorname{supp}(v) \subseteq \mathcal{Z}$. Choose $\epsilon > 0$ such that $|\epsilon v(x)| < |P(x)|$ for all $x \in \mathcal{Z}$. Then $P = \frac{1}{2}((P + \epsilon v) + (P - \epsilon v))$, where $P + \epsilon v, P - \epsilon v \in \mathcal{N}_P$. Therefore $v = 0$. The second statement follows from $\dim \mathbf{M}_A = \dim \mathcal{E}$. □

Lemma 3.6. *Let $\mathcal{Z} \subseteq \mathcal{X}$, and let $P \notin \overline{\mathcal{E}}$ be a local maximizer of $D_{\mathcal{E}}$ under the constraint that $\operatorname{supp}(P) \subseteq \mathcal{Z}$. If $\operatorname{supp}(P_{\mathcal{E}}) = \mathcal{X}$, then P is a local maximizer of $D_{\mathcal{E}}$.*

Proof. Let A be a sufficient statistics of $\mathcal{E}$. The composition of $\theta \mapsto P_\theta$ with the moment map π_A is continuously differentiable and injective (see Theorem 2.16). By assumption the image of the composition $\theta \mapsto \pi_A(P_\theta)$ contains a neighbourhood of $\pi_A(P_\mathcal{E})$. Since the differential has full rank (see the proof of Lemma 5.12 in Section 5.3 for a proof of this well-known fact), the inverse function theorem states that the inverse mapping $a \mapsto \theta(a)$ is continuously differentiable in a neighbourhood of $\pi_A(P_\mathcal{E})$. Moreover, since $P_\mathcal{E}$ has full support, the map $Q \mapsto D(Q\|\nu)$ is continuously differentiable in a neighbourhood of $P_\mathcal{E}$. Therefore, there is a neighbourhood U of P and a constant $C > 0$ such that $|D(Q_\mathcal{E}\|\nu) - D(Q'_\mathcal{E}\|\nu)| < C\|Q - Q'\|_\infty$ for all $Q, Q' \in U$. Let D be the maximum of $D(\cdot\|\nu)$ on $\mathbf{P}(\mathcal{X})$. Making U smaller if necessary one may assume that U satisfies the following three conditions:

(*i*) $Q \in U$ implies $Q^\mathcal{Z} \in U$.

(*ii*) P maximizes $D(Q\|\mathcal{E})$ subject to $Q \in U^\mathcal{Z} := U \cap \mathbb{R}^\mathcal{Z}$.

(*iii*) $h(s(Q), 1 - s(Q)) \geq (C + D)s(Q)$ for all $Q \in U$, where $s(Q) = Q(\mathcal{X} \setminus \mathcal{Z})$ and $h(s, 1 - s) = -s\log(s) - (1 - s)\log(1 - s)$ is the entropy of a binary random variable.

Fix $Q \in U$. If $\mathrm{supp}(Q) \subseteq \mathcal{Z}$, then $D(Q\|\mathcal{E}) \leq D(P\|\mathcal{E})$ by assumption. Otherwise write $Q = (1 - s)Q' + sR$, where $s = Q(\mathcal{X} \setminus \mathcal{Z}) > 0$, $Q' = Q^\mathcal{Z}$ and $R = Q^{\mathcal{X}\setminus\mathcal{Z}}$. Then $D(Q\|\nu) = (1 - s)D(Q'\|\nu) + sD(R\|\nu) - h(s, 1 - s)$. Theorem 2.16 (*ii*) implies

$$\begin{aligned} D(Q\|\mathcal{E}) &= D(Q\|\nu) - D(Q_\mathcal{E}\|\nu) \\ &= -h(s, 1-s) + s(D(R\|\nu) - D(Q'\|\nu)) \\ &\qquad + (D(Q'\|\nu) - D(Q'_\mathcal{E}\|\nu) + (D(Q'_\mathcal{E}\|\nu) - D(Q_\mathcal{E}\|\nu)) \\ &\leq -h(s, 1-s) + s(D(R\|\nu) - D(Q'\|\nu)) + D(Q'\|\mathcal{E}) + C\|Q - Q'\|_\infty \end{aligned}$$

Note that

$$\|Q - Q'\|_\infty = \max\left\{ s\max_{x\in\mathcal{Z}}\{R(x)\}, s\max_{x\in\mathcal{X}\setminus\mathcal{Z}} Q'(x)\} \right\} \leq s,$$

whence $D(Q\|\mathcal{E}) \leq D(P\|\mathcal{E}) - h(s, 1 - s) + (C + D)s \leq D(P\|\mathcal{E})$. □

Lemma 3.6 implies, in particular, that any point measure δ_x that rI-projects into $\mathcal{E}$ is a local maximizer. This statement is false in general if δ_x rI-projects into the boundary of $\mathcal{E}$, even under the additional assumption that $\delta_x \notin \overline{\mathcal{E}}$. One reason is Theorem 3.2. See 4.1.2 and 4.1.3 for counter-examples.

3.2. Projection points and kernel distributions

In this section assume that $\mathcal{E} \neq \mathbf{P}(\mathcal{X})^\circ$. Otherwise the function $D_\mathcal{E}$ is trivial, and there are no projection points. Under this assumption consider the map

$$\begin{aligned} \Psi_\mathcal{E} : \mathbf{P}(\mathcal{X}) \setminus \overline{\mathcal{E}} &\to \mathcal{N}, \\ P &\mapsto \frac{P - P_\mathcal{E}}{(P - P_\mathcal{E})^+(\mathcal{X})}. \end{aligned}$$

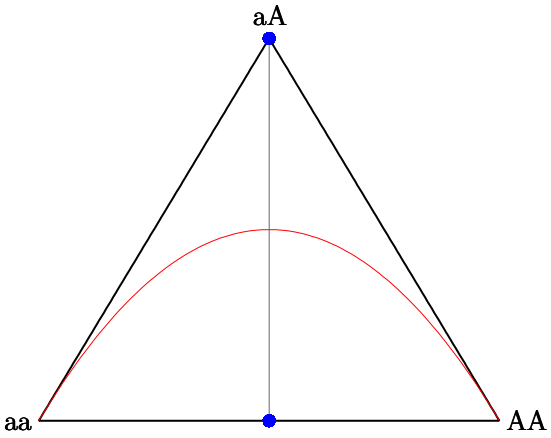

Figure 3.1.: A pair of projection points in the Hardy-Weinberg exponential family ($\nu(\mathrm{aa}) = \nu(\mathrm{AA}) = 1$, $\nu(\mathrm{Aa}) = 2$, $A = (0, 1, 2)$), see Section 4.4. The blue dots mark the two global maximizers, which are a pair of associated kernel distributions. The grey line connecting the two maximizers intersects the exponential family (in red) at their common rI-projection.

It maps each probability measure $P \in \mathbf{P}(\mathcal{X}) \setminus \overline{\mathcal{E}}$ to a vector which points from $P_{\mathcal{E}}$ to P. The normalization is chosen such that the image $u = \Psi_{\mathcal{E}}(P)$ is a difference of two probability distributions u^+, u^- of disjoint supports.

Lemma 3.7. *$P \in \mathbf{P}(\mathcal{X}) \setminus \overline{\mathcal{E}}$ is a projection point if and only if $\Psi_{\mathcal{E}}(P)^+ = P$.*

Proof. Let P be a projection point with support $\mathcal{Z}$. Let $P^- := \lambda P + (1-\lambda)P_{\mathcal{E}}$, where $\lambda = -\frac{P_{\mathcal{E}}(\mathcal{Z})}{1-P_{\mathcal{E}}(\mathcal{Z})}$. Then $P - P^- \in \mathcal{N}$, and

$$P^-(x) \begin{cases} 0, & \text{if } x \in \mathcal{Z}, \\ \frac{1+P_{\mathcal{E}}(\mathcal{Z})}{1-P_{\mathcal{E}}(\mathcal{Z})} P_{\mathcal{E}}(x) \geq 0, & \text{else.} \end{cases}$$

Hence P^- is a probability measure with support contained in $\mathcal{X} \setminus \mathcal{Z}$, and $u := P - P^-$ lies in the normal space $\mathcal{N}$. Geometrically, P^- is the point where the line through P in the direction of $P_{\mathcal{E}}$ leaves $\mathbf{P}(\mathcal{X})$. Both u and $\Psi_{\mathcal{E}}(P)$ are positive multiples of $P - P_{\mathcal{E}}$. Both are differences of probability distributions of disjoint supports, and hence they are equal, whence $P = u^+ = \Psi_{\mathcal{E}}(P)^+$.

Conversely, if $P = \Psi_{\mathcal{E}}(P)^+$, then the line through P in the direction $\Psi_{\mathcal{E}}(P)$ leaves $\mathbf{P}(\mathcal{X})$ in the measure $P^- := \Psi_{\mathcal{E}}(P)^-$. The rI-projection $P_{\mathcal{E}}$ must be on this line. Therefore, $P_{\mathcal{E}}$ is a convex combination of P and P^-. Since P and P^- have disjoint supports, P is a projection point. □

If P is a projection point, then it is easy to see that $\Psi_{\mathcal{E}}(P)^-$ is a second projection point with the same projection $P_{\mathcal{E}}$ to $\mathcal{E}$ as P. Figure 3.1 illustrates the situation.

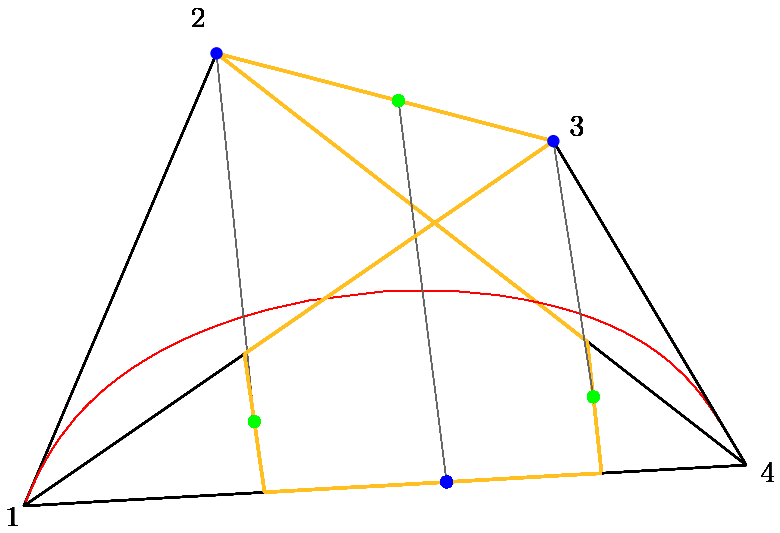

Figure 3.2.: An exponential family on four states (see Example 3.10). The blue dots are the local maximizers, green dots are the other projection points. Associated projection points are joined by grey lines. The set $\mathrm{K}_{\mathcal{E}}$ is marked in yellow.

Definition 3.8. Let $\mathcal{E}$ be an exponential family. A probability distribution $P \in \mathbf{P}(\mathcal{X})$ is called a *kernel distribution* of $\mathcal{E}$ if there exists a probability distribution $Q \in \mathbf{P}(\mathcal{X})$ such that $\operatorname{supp}(P) \cap \operatorname{supp}(Q) = \emptyset$ and $P - Q \in \mathcal{N}$. In this case P and Q are called *associated kernel distributions*. The set of all kernel distributions of $\mathcal{E}$ is denoted by $\mathrm{K}_{\mathcal{E}}$.

To any kernel distribution P there may be more than one associated kernel distribution $Q \in \mathbf{P}(\mathcal{X} \setminus \operatorname{supp}(P))$. See Remark 3.27 below for a natural choice of Q. Lemma 3.7 implies:

Corollary 3.9. *Every local maximizer of $D_{\mathcal{E}}$ is a kernel distribution*

As a consequence, the local maximizer of $D_{\mathcal{E}}$ can be found by maximizing the function $D_{\mathcal{E}}$ over $\mathrm{K}_{\mathcal{E}}$. This motivates the following mathematical problem:

- Find a "nice" statistical model $\mathcal{M}$ such that $\mathrm{K}_{\mathcal{E}}$ is contained in the closure of $\mathcal{M}$.

Ideally $\mathcal{M}$ should be a low-dimensional manifold in $\mathbf{P}(\mathcal{X})$ such that $\overline{\mathcal{M}} \setminus \mathcal{M} = \mathrm{K}_{\mathcal{E}}$. It is relatively easy to find such manifolds theoretically; for example, one could try to define $\mathcal{M}$ to be a minimal submanifold of $\mathbf{P}(\mathcal{X})$ with boundary $\mathrm{K}_{\mathcal{E}}$. To be useful for applications (as in Section 5.1) it is important to find a nice parametrization of $\mathcal{M}$. The following example shows that in general there is no exponential family with boundary $\mathrm{K}_{\mathcal{E}}$, since in general $\mathrm{K}_{\mathcal{E}}$ is not a union of exponential families.

Example 3.10. Figure 3.2 shows $\mathrm{K}_{\mathcal{E}}$ for the one-dimensional exponential family on four states with uniform reference measure and sufficient statistics $A = (0, 5, 12, 15)$. One-dimensional exponential families will be studied in more detail in Section 4.1.

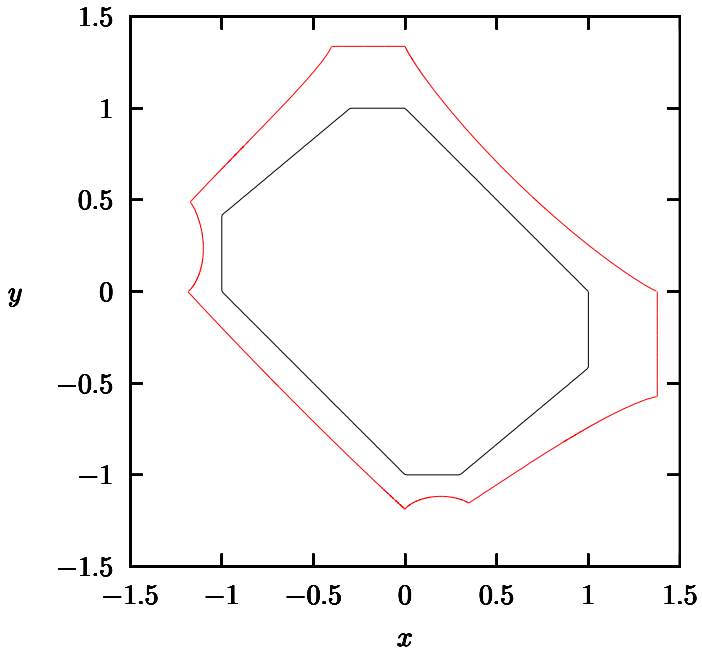

Figure 3.3.: A polar plot of the function $D_{\mathcal{E}}$. The black line is $\partial \mathbf{U}_{\mathcal{N}}$. See Example 3.11 for an explanation.

$\mathrm{K}_{\mathcal{E}}$ can be parametrized using the following construction: Let $u \in \mathcal{N} \setminus \{0\}$. Then $u = u^+ - u^-$, where u^+ and u^- are positive vectors of disjoint supports. By definition of $\mathcal{N}$ it follows that $d_u := u^+(\mathcal{X}) = u^-(\mathcal{X})$. Thus

$$u = d_u(P^+ - P^-),$$

where P^+ and P^- are two associated kernel distributions. The mapping $u \mapsto \frac{1}{d_u}u^+$ is a surjective parametrization $\mathcal{N} \setminus \{0\} \to \mathrm{K}_{\mathcal{E}}$. Therefore, the maximizers of $D_{\mathcal{E}}$ can be found by studying the function

$$u \in \mathcal{N} \setminus \{0\} \mapsto D(\frac{1}{d_u}u^+ \| \mathcal{E}).$$

It suffices to consider this function on the set $\{u \in \mathcal{N} : d_u = 1\}$. This set equals the boundary $\partial \mathbf{U}_{\mathcal{N}}$ of the polytope

$$\mathbf{U}_{\mathcal{N}} := \{u \in \mathcal{N} : d_u \leq 1\}.$$

$\partial \mathbf{U}_{\mathcal{N}}$ is the intersection of $\mathcal{N}$ with the sphere of radius 2 with respect to the ℓ_1-norm in $\mathbb{R}^{\mathcal{X}}$.

Example 3.11. Figure 3.3 gives a polar plot of the function $D_{\mathcal{E}}$ for the exponential family from Example 3.10. This plot was obtained by the following method: The normal space has dimension two, so it can be parametrized by two-dimensional polar

coordinates (here, the basis of 4.1.3 was used). Denote the unit vector with angular coordinate ϕ by u_ϕ. Then the point $\frac{1}{d_{u_\phi}} u_\phi$ lies in $\partial \mathbf{U}_\mathcal{N}$. To visualize the behaviour of $D(\frac{1}{d_{u_\phi}} u_\phi^+ \| \mathcal{E})$, the set

$$\left\{ \frac{1}{d_{u_\phi}} \left(1 + 0.3 \cdot D(\frac{1}{d_{u_\phi}} u_\phi^+ \| \mathcal{E}) \right) u_\phi \ : \ 0 \leq \phi < 2\pi \right\}$$

is plotted in red. This representation is chosen such that in regions, where $D_\mathcal{E}$ is constant, the red line is parallel to the boundary of the polytope, and the larger $D_\mathcal{E}$ is, the further the red line moves away from the black polytope (normalized by the distance of the polytope from the origin). The point measure δ_2 corresponds to the straight line at $x = 1$. On the corresponding face of $\mathbf{U}_\mathcal{N}$ all points u have the same positive part $u^+ = \delta_2$, and therefore, the function $D_\mathcal{E}$ is constant. Similarly, δ_3 corresponds to the face of $\mathbf{U}_\mathcal{N}$ given by $y = 1$. The third local maximum of $D_\mathcal{E}$ corresponds to a point on the line $x + y = -1$. This third maximum is inconspicuous and less pronounced than the other two maxima.

Some properties of the set $\mathrm{K}_\mathcal{E}$ are collected in the following proposition:

Proposition 3.12. *Let $\mathrm{K}_\mathcal{E}$ be the set of kernel distributions of the exponential family $\mathcal{E}$.*

(*i*) *Let $\mathcal{Y} \subset \mathcal{X}$. If $\mathbf{P}(\mathcal{Y})^\circ$ contains a kernel distribution, then $\mathrm{K}_\mathcal{E} \cap \mathbf{P}(\mathcal{Y})^\circ$ is convex, and the closure $\overline{\mathrm{K}_\mathcal{E} \cap \mathbf{P}(\mathcal{Y})^\circ}$ is a polytope. Hence $\mathrm{K}_\mathcal{E}$ is a union of polytopes.*

(*ii*) *If $\mathcal{Y} \subseteq \mathcal{X}$ is facial, then $\mathbf{P}(\mathcal{Y})^\circ \cap \mathrm{K}_\mathcal{E} = \emptyset$. In particular, $\mathrm{K}_\mathcal{E} \cap \overline{\mathcal{E}} = \emptyset$.*

(*iii*) *If $\mathcal{Z} \subset \mathcal{X}$ is not facial, then there is a kernel distribution P with $\mathrm{supp}(P) \subseteq \mathcal{Z}$.*

Proof. (*i*) If $u, v \in \partial \mathbf{U}_\mathcal{N}$ satisfy $\mathrm{supp}(u^+) = \mathrm{supp}(v^+)$, then $(1-\lambda)u^+ + \lambda v^+$ is the positive part of $((1-\lambda)u + \lambda v)^+$ for all $0 < \lambda < 1$. Since $\mathrm{K}_\mathcal{E} \cap \mathbf{P}(\mathcal{Y})^\circ = \{u^+ : u \in \mathbf{U}_\mathcal{N}, \mathrm{supp}(u^+) = \mathcal{Y}\}$ the closure equals the orthogonal projection of the polytope

$$\{u \in \mathbf{U}_\mathcal{N} : u(x) \geq 0 \text{ for all } x \in \mathcal{Y}, u(x) \leq 0 \text{ for all } x \notin \mathcal{Y}\}$$

to $\mathbb{R}^\mathcal{Y} \subseteq \mathbb{R}^\mathcal{X}$, and projections of polytopes are again polytopes (see Appendix A.1).

(*ii*) Let P, Q be associated kernel distributions. By Lemma 2.30, if $\mathcal{Y}$ is facial, then $\mathrm{supp}(P) \subseteq \mathcal{Y}$ implies $\mathrm{supp}(Q) \subseteq \mathcal{Y}$, and so $P \notin \mathbf{P}(\mathcal{Y})^\circ$.

(*iii*) Let A be a sufficient statistics of $\mathcal{E}$, and let $\mathbf{F}$ be the smallest face of $\mathbf{M}_A$ containing $\{A_x : x \in \mathcal{Z}\}$. By assumption there exists $x \in \mathcal{X} \setminus \mathcal{Z}$ such that $A_x \in \mathbf{F}$. Therefore, the set $\{A_y : y \in \mathcal{Z} \cup \{x\}\}$ is affinely dependent, so there exists a $v \in \mathcal{N}$ such that $\mathrm{supp}(v) \subseteq \mathcal{Z} \cup \{x\}$. Without loss of generality assume that $v(x) \leq 0$, then $P := \frac{1}{|v^+(\mathcal{X})|} v^+$ is a kernel distribution in $\mathbf{P}(\mathcal{Z})$. □

Example 3.13. If $\mathcal{N}$ is one-dimensional, then $\partial \mathbf{U}_\mathcal{N}$ consists of only two points $\pm u$, where $u = P^+ - P^-$ is a difference of two associated kernel distributions P^+, P^-. One of P^+ and P^- must be a global maximizer. It may happen that they are both

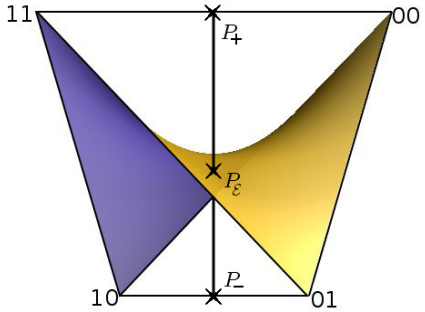

Figure 3.4.: The independence model of two binary variables.

global maximizers, see Figure 3.1. In any case, both P^+ and P^- are local maximizers. This fact will follow easily from Theorem 3.28 below. Alternatively, this can be seen directly as follows: By Theorem 2.21, $\mathcal{E}$ equals the solution set of the equation $P^{P^+} - P^{P^-} = 0$; hence the maximum of $D_{\mathcal{E}}$ among all probability distributions P such that $P^{P^+} - P^{P^-} \geq 0$ is at P^+, and the maximum of $D_{\mathcal{E}}$ among all probability distributions P such that $P^{P^+} - P^{P^-} \leq 0$ is at P^-.

This result can serve as a source of examples and counterexamples. For example, it is easy to see that any probability measure P^+ supported on a set $\mathcal{Y} \subset \mathcal{X}$ of cardinality less than $|\mathcal{X}|$ is a global maximizer of $D_{\mathcal{E}}$ for some exponential family $\mathcal{E}$: Just choose $P^- \in \mathbf{P}(\mathcal{X} \setminus \mathcal{Y})$ and let $\mathcal{E}$ be an exponential family with normal space $\mathcal{N}$ spanned by u. By selecting an appropriate reference measure, either P^+ or P^- can be made the global maximum. The same argument proves that the support of the rI-projection of a local maximizer can be an arbitrary set $\mathcal{Y}$ of cardinality at least two. Of course, when the reference measure is fixed or when the class of exponential families is restricted in any other way, these statements are not true in general anymore.

As a special case, consider the independence model of two binary units, see Figure 3.4. A sufficient statistics is

$$A_{2,2} = \begin{pmatrix} 00 & 01 & 10 & 11 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}.$$

By Theorem 2.48 the normal space is spanned by

$$u = (+1, -1, -1, +1),$$

corresponding to two global maximizers $P^+ = \frac{1}{2}(\delta_{00} + \delta_{11})$ and $P^- = \frac{1}{2}(\delta_{01} + \delta_{10})$.

3.3. The function $\overline{D}_{\mathcal{E}}$

Example 3.13 shows that Corollary 3.9 is a valuable tool to study the maximizers of $D_{\mathcal{E}}$. Yet it is still difficult to obtain the function $u \in \partial \mathbf{U}_{\mathcal{N}} \mapsto D_{\mathcal{E}}(u^+)$, because

it is difficult to compute rI-projections, see Remark 2.20. Fortunately, it is possible to replace this function by another function, which is easier to compute and more naturally defined directly on $\mathcal{N}$.

Let P^+ be a projection point with support $\mathcal{Z}$, and let $P^- = \Psi_{\mathcal{E}}(P^+)^-$. The rI-projection $P_{\mathcal{E}}$ of P^+ and P^- can be written as a convex combination of P^+ and P^-, i.e. $P_{\mathcal{E}} = \mu P^+ + (1-\mu)P^-$, where $0 < \mu < 1$. Since the supports of P^+ and P^- are disjoint, $\mu = P_{\mathcal{E}}(\mathcal{Z})$ and $(1-\mu) = P_{\mathcal{E}}(\mathcal{X} \setminus \mathcal{Z})$. In other words,

$$P_{\mathcal{E}}(x) = \begin{cases} \mu P^+(x), & x \in \mathcal{Z}, \\ (1-\mu)P^-(x), & x \notin \mathcal{Z}. \end{cases}$$

When P^+ and P^- are known, the rI-projection $P_{\mathcal{E}}$ can be characterized as the unique probability measure that minimizes the strictly convex function $D(\cdot\|\nu)$ over the convex hull of P^+ and P^-, see Theorem 2.16. The following lemma collects some consequences for similar triples of probability measures:

Lemma 3.14. *Let P^+ and P^- be two associated kernel distributions. Let $\hat{P}$ be the unique probability measure that minimizes $D(\cdot\|\nu)$ over the convex hull of P^+ and P^-. Define $\mu = \hat{P}(\mathcal{Z})$, where $\mathcal{Z} = \operatorname{supp}(P^+)$. Then the following equations hold:*

$$\exp(-D(\hat{P}\|\nu)) = \exp(-D(P^+\|\nu)) + \exp(-D(P^-\|\nu)), \tag{3.3a}$$

$$\frac{\mu}{1-\mu} = \exp\left(D(P^-\|\mathcal{E}) - D(P^+\|\mathcal{E})\right), \tag{3.3b}$$

$$\begin{aligned} D(P^+\|\hat{P}) &= D(P^+\|\nu) - D(\hat{P}\|\nu) \\ &= \log(1 + \exp(D(P^+\|\mathcal{E}) - D(P^-\|\mathcal{E}))). \end{aligned} \tag{3.3c}$$

Proof. The first observation is

$$D(\hat{P}\|\nu) = \mu D(P^+\|\nu) + (1-\mu)D(P^-\|\nu) - h(\mu, 1-\mu), \tag{3.4}$$

where $h(\mu, 1-\mu) = -\mu\log(\mu) - (1-\mu)\log(1-\mu)$. Since $\hat{P}$ minimizes $D(\cdot\|\nu)$ on the convex hull of P^+ and P^-, it follows that

$$\begin{aligned} &\frac{\partial\left(\mu' D(P^+\|\nu) + (1-\mu')D(P^-\|\nu) - h(\mu', 1-\mu')\right)}{\partial\mu'}\bigg|_{\mu'=\mu} \\ &\qquad = D(P^+\|\nu) - D(P^-\|\nu) - \log(1-\mu) + \log(\mu) \end{aligned}$$

vanishes, which rewrites to

$$\frac{\mu}{1-\mu} = \exp\left(D(P^-\|\nu) - D(P^+\|\nu)\right).$$

Theorem 2.16 (*ii*) shows $D(Q^+\|\nu) - D(Q^-\|\nu) = D(Q^+\|\mathcal{E}) - D(Q^-\|\mathcal{E})$ whenever $Q^+ - Q^- \in \mathcal{N}$, whence (3.3b). Solving for μ yields

$$\mu = \frac{\exp(-D(P^+\|\nu))}{\exp(-D(P^+\|\nu)) + \exp(-D(P^-\|\nu))} = \frac{1}{1 + \exp(D(P^+\|\nu) - D(P^-\|\nu))}. \tag{3.5}$$

This implies

$$h(\mu, 1-\mu) = \mu D(P^+\|\nu) + (1-\mu)D(P^-\|\nu) + \log\left(\exp(-D(P^+\|\nu)) + \exp(-D(P^-\|\nu))\right).$$

Comparison with (3.4) yields (3.3a), which in turn transforms (3.5) into

$$\mu = \exp(D(\hat{P}\|\nu) - D(P^+\|\nu)).$$

The information divergence equals

$$\begin{aligned} D(P^+\|\hat{P}) &= \sum_{x\in\mathcal{Z}} P^+(x) \log \frac{1}{\hat{P}(\mathcal{Z})} = -\log(\mu) \\ &= D(P^+\|\nu) - D(\hat{P}\|\nu) \\ &= \log(1 + \exp(D(P^+\|\mathcal{E}) - D(P^-\|\mathcal{E}))), \end{aligned}$$

completing the proof of the lemma. □

Remark 3.15. As an easy consequence

$$\exp(-D(P^+\|\mathcal{E})) + \exp(-D(P^-\|\mathcal{E})) = 1.$$

Hence, in general only one of P^+ and P^- is a local maximizer of $D_{\mathcal{E}}$. Furthermore, $D(P\|\mathcal{E}) \geq \log(2)$ for any global maximizer P (assuming that $\overline{\mathcal{E}} \neq \mathbf{P}(\mathcal{X})$).

These facts can be used to relate two different optimization problems. The first one is the maximization of $D_{\mathcal{E}}$. The second one is the maximization of the function

$$\overline{D}_{\mathcal{E}} : \partial\mathbf{U}_{\mathcal{N}} \to \mathbb{R}, \qquad u \mapsto \sum_x u(x) \log \frac{|u(x)|}{\nu_x}. \tag{3.6}$$

Theorem 2.16 (*ii*) shows that $\overline{D}_{\mathcal{E}}$ satisfies

$$\overline{D}_{\mathcal{E}}(u) = D(u^+\|\nu) - D(u^-\|\nu) = D(u^+\|\mathcal{E}) - D(u^-\|\mathcal{E}). \tag{3.7}$$

In particular, $\overline{D}_{\mathcal{E}}$ does not depend on the choice of the reference measure. Since $\overline{D}_{\mathcal{E}}$ is a continuous function on a compact set $\partial\mathbf{U}_{\mathcal{N}}$, a maximum is guaranteed to exist.

Theorem 3.16. *Let $\mathcal{E}$ be an exponential family with normal space $\mathcal{N}$. The map $\Psi_{\mathcal{E}}$ restricts to a bijection from the set of global maximizers of $D_{\mathcal{E}}$ to the set of global maximizers of $\overline{D}_{\mathcal{E}}$. An inverse is given by the restriction of the map $\Psi^+ : u \mapsto u^+$. If $P \in \mathbf{P}(\mathcal{X})$ and $u \in \partial\mathbf{U}_{\mathcal{N}}$ are global maximizers of $D_{\mathcal{E}}$ and $\overline{D}_{\mathcal{E}}$, respectively, then*

$$D(P\|\mathcal{E}) = \log(1 + \exp(\overline{D}_{\mathcal{E}}(u)))$$

Later, Theorem 3.28 will show that the projection points and the local maximizers of $D_{\mathcal{E}}$ can also be found by studying the function $\overline{D}_{\mathcal{E}}$. The proof of Theorem 3.16 builds upon the following lemma:

Lemma 3.17. *$D(u^+\|\mathcal{E}) \geq \log(1+\exp(\overline{D}_\mathcal{E}(u)))$ for all $u \in \partial\mathbf{U}_\mathcal{N}$, with equality if and only if u^+ is a projection point.*

Proof. Let $\mathcal{E}'$ be the exponential family with normal space $\mathcal{N}' = \mathbb{R}u$ and reference measure ν. Then u^+ is a projection point of $\mathcal{E}'$. Furthermore, $\mathcal{E} \subseteq \mathcal{E}'$. By Lemma 3.14,

$$\log(1+\exp(\overline{D}_\mathcal{E}(u))) = \log(1+\exp(\overline{D}_{\mathcal{E}'}(u))) = D(u^+\|\mathcal{E}') \leq D(u^+\|\mathcal{E}).$$

Equality holds if and only if $(u^+)_{\mathcal{E}'} = (u^+)_\mathcal{E}$. In this case $(u^+)_\mathcal{E}$ lies in the convex hull of u^+ and u^-, which means that u^+ is a projection point. □

Proof of Theorem 3.16. If $u \in \partial\mathbf{U}_\mathcal{N}$ is a global maximizer of $\overline{D}_\mathcal{E}$ and if $P \in \mathbf{P}(\mathcal{X})$ is a global maximizer of $D_\mathcal{E}$, then

$$D_\mathcal{E}(P) = D_\mathcal{E}(\Psi_\mathcal{E}(P)^+) = \log(1+\exp(\overline{D}_\mathcal{E}(\Psi_\mathcal{E}(P)))) \leq \log(1+\exp(\overline{D}_\mathcal{E}(u))) \leq D_\mathcal{E}(u^+)$$

by Lemmas 3.7 and 3.17. By the maximality assumptions these inequalities hold as equalities, so $\overline{D}_\mathcal{E}(\Psi_\mathcal{E}(P)) = \overline{D}_\mathcal{E}(u)$ and $D_\mathcal{E}(u^+) = D_\mathcal{E}(P)$. It remains to show that the restrictions of $\Psi_\mathcal{E}$ and Ψ^+ are mutually inverse bijections. Lemma 3.7 implies that $\Psi_\mathcal{E}$ is injective and Ψ^+ is surjective on the global maximizers; hence it suffices to show that Ψ^+ is injective on the global maximizers of $\overline{D}_\mathcal{E}$. Let $u, v \in \partial\mathbf{U}_\mathcal{N}$ be two global maximizers of $\overline{D}_\mathcal{E}$ such that $u^+ = v^+$. By (3.7) both u^- and v^- minimize the convex function $D(\cdot\|\nu)$ on $\mathbf{P}(\mathcal{X} \setminus \operatorname{supp}(u^+))$, whence $u^- = v^-$. □

Remark 3.18. $\overline{D}_\mathcal{E}$ can be extended to a function on $\mathcal{N}$ using (3.6) or, equivalently, (3.7). Then $\overline{D}_\mathcal{E}$ is homogeneous of degree one on $\mathcal{N}$: For any $\alpha \in \mathbb{R}$

$$\overline{D}_\mathcal{E}(\alpha u) = \alpha \sum_x u(x) \log \frac{|u(x)|}{\nu_x} + \alpha \left(\sum_x u(x) \right) \log|\alpha| = \alpha \overline{D}_\mathcal{E}(u).$$

Hence the global maximizers of $\overline{D}_\mathcal{E}$ on $\mathbf{U}_\mathcal{N}$ agree with the global maximizers on $\partial\mathbf{U}_\mathcal{N}$. Another possibility is to maximize the function

$$\overline{D}^1_\mathcal{E} : \mathcal{N} \setminus \{0\} \to \mathbb{R}, \quad u \mapsto \overline{D}_\mathcal{E}(u/d_u) = \frac{1}{d_u} \overline{D}_\mathcal{E}(u),$$

where $d_u = u^+(\mathcal{X})$. A signed measure $u \in \mathcal{N} \setminus \{0\}$ is a local maximizer of $\overline{D}^1_\mathcal{E}$ if and only if $\frac{1}{d_u}u$ is a local maximizer of $\overline{D}_\mathcal{E}$ on $\partial\mathbf{U}_\mathcal{N}$.

Example 3.19. Figure 3.5 gives a polar plot of $D_\mathcal{E}$ and $\overline{D}_\mathcal{E}$ for the exponential family from Example 3.10. The polar plot was obtained in the same way as in Example 3.11. The one-to-one correspondence between the local maximizers, which will be proved in Theorem 3.28, is also clearly visible.

Example 3.20. Consider again the case where $\dim\mathcal{N} = 1$, which was discussed already in Example 3.13. The value of $D_\mathcal{E}$ at the two local maximizers P^+ and P^- is easy to compute in terms of $\overline{D}_\mathcal{E}(u)$: By Lemma 3.14

$$D(P^+\|\mathcal{E}) = \log(1+\exp(\overline{D}_\mathcal{E}(u))), \qquad D(P^-\|\mathcal{E}) = \log(1+\exp(-\overline{D}_\mathcal{E}(u))).$$

In particular, if $\mathbf{1}$ is a reference measure of $\mathcal{E}$, then P^+ is a global maximum of $D_\mathcal{E}$ if and only if the entropy of P^+ is not greater than the entropy of P^-.

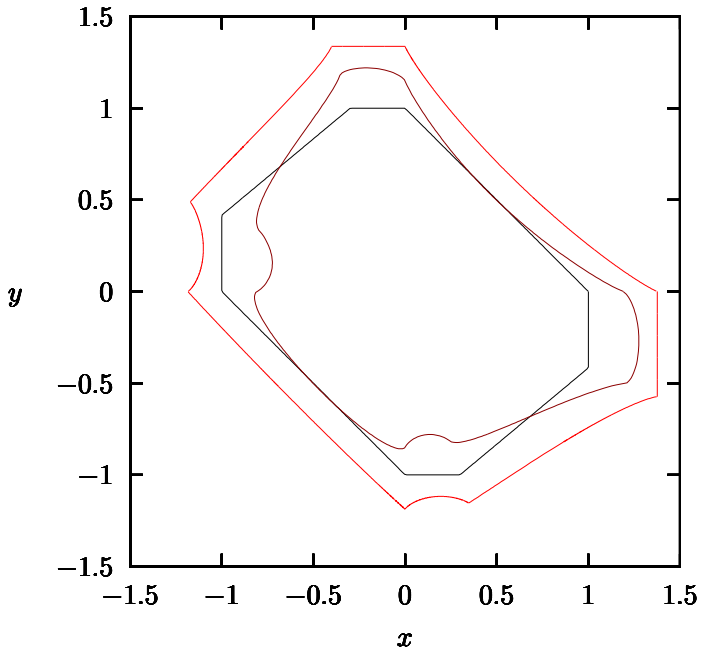

Figure 3.5.: A polar plot of $D_{\mathcal{E}}$ and $\overline{D}_{\mathcal{E}}$ for Example 3.19.

3.4. The first order optimality conditions of $\overline{D}_{\mathcal{E}}$

The goal of this section is to compute the first order optimality conditions of $\overline{D}_{\mathcal{E}}$. The next section will show that the bijection of Theorem 3.16 between the global maximizers of $D_{\mathcal{E}}$ and $\overline{D}_{\mathcal{E}}$ extends to a bijection of all critical points.

Proposition 3.21. *Let $\mathcal{E}$ be an exponential family with normal space $\mathcal{N}$, let $u \in \partial \mathbf{U}_{\mathcal{N}}$ be a local maximizer of $\overline{D}_{\mathcal{E}}$, and let $\mathcal{Y} = \operatorname{supp}(u)$. The following statements hold:*

(*i*) *$v(\mathcal{Y}) = 0$ for all $v \in \mathcal{N}$.*

(*ii*) *If $v \in \mathcal{N}$ satisfies $\operatorname{supp}(v) \subseteq \mathcal{Y}$ and $v(u > 0) = 0$, then*

$$\sum_{x:u\neq 0} v(x) \log \frac{|u(x)|}{\nu_x} = 0. \tag{3.8}$$

(*iii*) *Let $P_{\mathcal{E}}$ be the rI-projection of u^+ and u^-. If $\mathcal{Y} \neq \mathcal{X}$, then*

$$\overline{D}_{\mathcal{E}^{P_{\mathcal{E}}}}(v) \leq \overline{D}_{\mathcal{E}}(u), \qquad \text{for all } v \in \mathbf{U}_{\mathcal{N}'}, \tag{3.9}$$

where $\mathcal{N}' = \{v|_{\mathcal{X}\setminus\mathcal{Y}} : v \in \mathcal{N}\}$ is the normal space of the exponential family $\mathcal{E}^{P_{\mathcal{E}}} = \{Q^{\mathcal{X}\setminus\mathcal{Y}} : Q \in \mathcal{E} \text{ and } Q^{\mathcal{Y}} = P_{\mathcal{E}}\}$.

These are all first-order optimality conditions.

Definition 3.22. A point $u \in \mathcal{N} \setminus \{0\}$ is called a *facial difference of projection points (f.d.p.)* if $\operatorname{supp}(u)$ is facial and if u satisfies equations (3.8) for all $v \in \mathcal{N}$ satisfying $\operatorname{supp}(v) \subseteq \operatorname{supp}(u)$ and $v(u > 0) = 0$. If $u \in \mathcal{N} \setminus \{0\}$ satisfies statements (i) and (ii) of Proposition 3.21, then it is a *quasi-critical point* of $\overline{D}_{\mathcal{E}}$, and $u \in \mathcal{N} \setminus \{0\}$ is called a *critical point* if it satisfies all three statements of Proposition 3.21.

The name "f.d.p." is justified by Lemma 3.25 below.

Remark 3.23. Only critical points which lie in $\partial \mathbf{U}_{\mathcal{N}}$ can be local maximizers of $\overline{D}_{\mathcal{E}}$. If u is an arbitrary critical point, then $\frac{1}{d_u} u \in \partial \mathbf{U}_{\mathcal{N}}$ is also a critical point, cf. Remark 3.18. Similar remarks hold for the quasi-critical points and f.d.p.s.

The proof of the proposition makes use of the following two lemmas. The first shows that condition (i) can be seen as an algebraic version of statement (i) of Theorem 3.2. While the geometric notion of parallel hyperplanes can be quickly read off on low dimensional convex supports, the algebraic version is easier to check by a computer.

Lemma 3.24. *Let $\mathcal{Y} \subsetneq \mathcal{X}$, let $A \in \mathbb{R}^{h \times \mathcal{X}}$, and let $\mathcal{N} = \{u \in \ker A : \sum_x u(x) = 0\}$. Then the following two conditions are equivalent:*

(i) *The sets $\{A_x : x \in \mathcal{Y}\}$ and $\{A_x : x \notin \mathcal{Y}\}$ lie in distinct parallel hyperplanes.*

(ii) *All $u \in \mathcal{N}$ satisfy $u(\mathcal{Y}) = 0$.*

Proof. If (i) holds, then there exist real numbers $a \neq b$ and a vector $c \in \mathbb{R}^h$ such that $\sum_{i=1}^h c_i A_{i,x} = a$ for all $x \in \mathcal{Y}$ and $\sum_{i=1}^h c_i A_{i,x} = b$ for all $x \in \mathcal{X} \setminus \mathcal{Y}$. Let $u \in \mathcal{N}$. Then

$$0 = \sum_{i=1}^{h} c_i \sum_{x \in \mathcal{X}} A_{i,x} u(x) = a u(\mathcal{Y}) + b u(\mathcal{X} \setminus \mathcal{Y}).$$

Together with $0 = u(\mathcal{X}) = u(\mathcal{Y}) + u(\mathcal{X} \setminus \mathcal{Y})$ this implies (ii).

Conversely, if (i) does not hold, then the two affine spaces generated by $\{A_x\}_{x \in \mathcal{Y}}$ and $\{A_x\}_{x \in \mathcal{X} \setminus \mathcal{Y}}$ must meet in a nontrivial vector $z \in \mathbb{R}^h \setminus \{0\}$. Write $z = \sum_{x \in \mathcal{Y}} \alpha_x A_x = \sum_{x \in \mathcal{X} \setminus \mathcal{Y}} \beta_x A_x$, where $\sum_{x \in \mathcal{Y}} \alpha_x = 1 = \sum_{x \in \mathcal{X} \setminus \mathcal{Y}} \beta_x$. Define $v \in \mathbb{R}^{\mathcal{X}}$ via

$$v(x) = \begin{cases} \alpha_x, & \text{if } x \in \mathcal{Y}, \\ -\beta_x, & \text{else.} \end{cases}$$

Then $Av = z - z = 0$, so $v \in \mathcal{N}$, and $\sum_{x \in \mathcal{Y}} v(x) = 1$, so (ii) holds neither. □

Lemma 3.25. *Let $u \in \partial \mathbf{U}_{\mathcal{N}}$, and let $P_{\mathcal{E}}$ be the rI-projection of u^+ and u^-. Then u is a f.d.p. if and only if $P_{\mathcal{E}}$ is contained in the convex hull of u^+ and u^-. In particular, if u is a f.d.p., then u^+ and u^- are projection points, and* $\operatorname{supp}(u)$ *is facial.*

Proof. Let $u \in \partial \mathbf{U}_{\mathcal{N}}$, and let $\hat{P}$ be the probability measure that minimizes $D_{\mathcal{E}}$ on the convex hull of u^+ and u^-. For any $v \in \mathcal{N}$ such that $\operatorname{supp}(v) \subseteq \operatorname{supp}(\hat{P})$ the directional

derivative of $D(\cdot\|\nu)$ in the direction v at $\hat{P}$ equals

$$\sum_{x\in\mathrm{supp}(v)} v(x)\log\frac{\hat{P}(x)}{\nu_x} = \sum_{x\in\mathrm{supp}(v)} v(x)\log\frac{|u(x)|}{\nu_x} + v(u>0)\log\mu + v(u<0)\log(1-\mu)$$
$$= \sum_{x\in\mathrm{supp}(v)} v(x)\log\frac{|u(x)|}{\nu_x} - v(u>0)\overline{D}_{\mathcal{E}}(u), \tag{3.10}$$

using $\mu = \hat{P}(u>0)$ and Lemma 3.14.

If u is a f.d.p., then $\mathrm{supp}(u)$ is facial, and so $\mathrm{supp}(P_{\mathcal{E}}) \subseteq \mathrm{supp}(u)$ by Lemma 2.28. By (ii) the directional derivative (3.10) vanishes for $v = P_{\mathcal{E}} - \hat{P}$. But on the convex hull of $\hat{P}$ and $P_{\mathcal{E}}$ the function $D_{\mathcal{E}}$ is strictly convex and attains its minimum at $P_{\mathcal{E}}$, whence $\hat{P} = P_{\mathcal{E}}$.

Conversely, if $P_{\mathcal{E}} = \hat{P}$, then (3.10) vanishes for all $v \in \mathcal{N}$ such that $\mathrm{supp}(v) \subseteq \mathrm{supp}(u)$. Furthermore, $\mathrm{supp}(u) = \mathrm{supp}(P_{\mathcal{E}})$ is facial, and so u is a f.d.p. □

Proof of Proposition 3.21. The degree $d_v = \sum_x v^+(x) = \sum_x v^-(x)$ is piecewise linear in the following sense:

- Let $u, v \in \mathcal{N}$. Then there exists $\lambda_1 > 0$ such that

$$d_{u+\lambda v} = d_u + \lambda d'_u(v) \text{ for all } 0 \le \lambda \le \lambda_1, \tag{3.11}$$

  where $d'_u(v) = \sum_{x:u>0} v(x) + \sum_{x:u=0} v^+(x) = v(u>0) + v^+(u=0) \in \mathbb{R}$ depends only on u and v (but not on λ).

Fix $u, v \in \mathcal{N}$. If $\epsilon > 0$ is small enough, then

$$\overline{D}_{\mathcal{E}}(u+\epsilon v) = \sum_x u(x)\log\frac{|u(x)|}{\nu_x}$$
$$+ \sum_{x:u\neq 0} u(x)\log\left(1+\epsilon\frac{v(x)}{u(x)}\right) + \epsilon\sum_x v(x)\log\frac{|u(x)+\epsilon v(x)|}{\nu_x}$$
$$= \overline{D}_{\mathcal{E}}(u) + \epsilon\left(\sum_{x:u\neq 0} v(x)\log\frac{|u(x)|}{\nu_x} + \sum_{x:u=0} v(x)\log\frac{|v(x)|}{\nu_x}\right)$$
$$+ \epsilon\log|\epsilon| v(u=0) + \epsilon v(u\neq 0) + o(\epsilon),$$

where $\log(1+\epsilon x) = \epsilon x + o(\epsilon)$ was used. Equation (3.11) yields

$$\overline{D}^1_{\mathcal{E}}(u+\epsilon v) = \overline{D}^1_{\mathcal{E}}(u) - \epsilon\frac{d'_u(v)}{d_u}\overline{D}^1_{\mathcal{E}}(u)$$
$$+ \frac{\epsilon}{d_u}\left(\sum_{x:u\neq 0} v(x)\log\frac{|u(x)|}{\nu_x} + \sum_{x:u=0} v(x)\log\frac{|v(x)|}{\nu_x}\right)$$
$$+ \frac{1}{d_u}\left(\epsilon\log|\epsilon| v(u=0) + \epsilon v(u\neq 0)\right) + o(\epsilon)$$

for the function $\overline{D}^1_{\mathcal{E}}$ from Remark 3.18.

Let $u \in \partial \mathbf{U}_{\mathcal{N}}$ be a local maximizer of $\overline{D}_{\mathcal{E}}$. Then u is also a local maximizer of $\overline{D}^1_{\mathcal{E}}$ by Remark 3.18. The derivative of $\epsilon \log \epsilon$ diverges at zero, and if v is replaced by $-v$, then the coefficient $\frac{1}{d_u} v(u = 0)$ changes its sign. Therefore, $v(u = 0) = 0$, and hence the first statement follows from $v(\mathcal{Y}) = v(\mathcal{X}) - v(u = 0) = 0$. This implies

$$\sum_{x:u\neq 0} v(x) \log \frac{|u(x)|}{\nu_x} + \sum_{x:u=0} v(x) \log \frac{|v(x)|}{\nu_x} \leq d'_u(v) \overline{D}_{\mathcal{E}}(u) \tag{3.12}$$

for all $v \in \mathcal{N}$. If $\operatorname{supp}(v) \subseteq \operatorname{supp}(u)$, then $d'_u(-v) = -v(u > 0) = -d'_u(v)$. In this case the two sides of the inequality change their sign when v is replaced by $-v$, and thus

$$\sum_{x:u\neq 0} v(x) \log \frac{|u(x)|}{\nu_x} = d'_u(v) \overline{D}_{\mathcal{E}}(u), \tag{3.13}$$

for all $v \in \mathcal{N}$ satisfying $\operatorname{supp}(v) \subseteq \operatorname{supp}(u)$. The proof now follows in two steps, which reformulate (3.12) and (3.13):

(1) *Condition (ii) is equivalent to equations* (3.13) *for all* $v \in \mathcal{N}$ *that satisfy* $\operatorname{supp}(v) \subseteq \operatorname{supp}(u)$*:* The function $d'_u(v)$ is linear on $\mathcal{N} \cap \mathbb{R}^{\mathcal{Y}}$. Hence (3.13) is linear in v and trivially satisfied for $v = u$. Therefore, it is enough to check (3.13) on a vector space complement of $\mathbb{R}u$ in $\mathcal{N} \cap \mathbb{R}^{\mathcal{Y}}$. Since $d'_u(u) = 1$ such a complement can be defined by $d'_u(v) = 0$.

(2) *Given (i) and (ii), condition (iii) is equivalent to the inequalities* (3.12) *for all* $v \in \mathcal{N}$*:* If $\mathcal{Y} = \mathcal{X}$, then there is nothing to show, so assume $\mathcal{Y} \neq \mathcal{X}$, and let $\mathcal{Y}' = \mathcal{X} \setminus \mathcal{Y}$. Suppose that $u \in \mathbf{U}_{\mathcal{N}}$ is quasi-critical. By Lemma 3.24 the support of u is facial. Hence u^+ and u^- are projection points by Lemma 3.25. Denote the rI-projection of u^+ and u^- by $P_{\mathcal{E}}$. By Theorem 2.29 there exists $Q \in \mathcal{E}$ such that $P_{\mathcal{E}} = Q^{\mathcal{Y}}$. Hence $Q|_{\mathcal{Y}'}$ is a reference measure for $\mathcal{E}^{P_{\mathcal{E}}}$. Let $v \in \mathcal{N}$. Then

$$\begin{aligned}
&\sum_{x\in\operatorname{supp}(u)} v(x) \log \frac{|u(x)|}{\nu_x} \\
&\qquad = \sum_{x\in\operatorname{supp}(u)} v(x) \log \frac{Q(x)}{\nu_x} - v(u > 0) \log Q(u > 0) - v(u < 0) \log Q(u < 0).
\end{aligned}$$

By definition of the normal space, v is orthogonal to the vector $\log \frac{Q}{\nu}$, so

$$\sum_{x\in\operatorname{supp}(u)} v(x) \log \frac{Q(x)}{\nu_x} = - \sum_{x\in\operatorname{supp}(v)\setminus\operatorname{supp}(u)} v(x) \log \frac{Q(x)}{\nu_x}.$$

Furthermore, $Q(u > 0) = P_{\mathcal{E}}(u > 0)Q(\mathcal{Y})$ and $Q(u < 0) = P_{\mathcal{E}}(u < 0)Q(\mathcal{Y})$. By

assumption $v(u>0)+v(u<0)=v(u\neq 0)=0$, so Lemma 3.14 implies

$$\sum_{x\in\mathrm{supp}(u)} v(x)\log\frac{|u(x)|}{\nu_x} = -\sum_{x\in\mathrm{supp}(v)\setminus\mathrm{supp}(u)} v(x)\log\frac{Q(x)}{\nu_x} - v(u>0)\log\frac{P_{\mathcal{E}}(u>0)}{P_{\mathcal{E}}(u<0)}.$$
$$= -\sum_{x\in\mathrm{supp}(v)\setminus\mathrm{supp}(u)} v(x)\log\frac{Q(x)}{\nu_x} + v(u>0)\overline{D}_{\mathcal{E}}(u),$$

Hence inequality (3.12) is equivalent to

$$\overline{D}_{\mathcal{E}^{P_{\mathcal{E}}}}(v|_{\mathcal{Y}'}) = \sum_{x\in\mathcal{Y}'} v(x)\log\frac{|v(x)|}{Q(x)} \leq v^+(u=0)\overline{D}_{\mathcal{E}}(u).$$

Therefore, the claim follows from Lemma 3.4, Remark 3.18 and the observation that

$$\left\{\frac{1}{v^+(\mathcal{Y}')}v|_{\mathcal{Y}'} : v\in\mathcal{N}\right\} = \mathbf{U}_{\mathcal{N}'}. \qquad \square$$

The next lemma shows that the dimension of the set of f.d.p.s is bounded from above by the dimension of the exponential family:

Lemma 3.26. *Let $\mathcal{E}$ be an exponential family with normal space $\mathcal{N}$, let $u\in\mathcal{N}$ be a f.d.p., and let $\mathcal{Z}=\mathcal{X}\setminus\mathrm{supp}(u^+)$. Denote by $\mathcal{E}^{\mathcal{Z}}$ the exponential family on $\mathcal{Z}$ with reference measure $\nu|_{\mathcal{Z}}$ and normal space $\{v|_{\mathcal{Z}} : v\in\mathcal{N}\}$. Then $u^-\in\overline{\mathcal{E}^{\mathcal{Z}}}$.*

Proof. Since $\mathrm{supp}(u)$ is facial for $\mathcal{E}$, it follows that $\mathrm{supp}(u^-)$ is facial for $\mathcal{E}^{\mathcal{Z}}$. By Lemma 3.25 and the assumptions u^- is a projection point, and the rI-projection $P_{\mathcal{E}}$ of u^- has support $\mathrm{supp}(u)$. By Theorem 2.29 there exists $Q\in\mathcal{E}$ such that $Q^{\mathrm{supp}(u)}=P_{\mathcal{E}}$. Therefore, $u^- = P_{\mathcal{E}}^{\mathrm{supp}(u^-)} = Q^{\mathrm{supp}(u^-)}\in\overline{\mathcal{E}^{\mathcal{Z}}}$ by Theorem 2.29. $\square$

Remark 3.27. Lemma 3.26 suggests to assign a unique associated kernel distribution $\Phi(P)$ to any given kernel distribution $P\in\mathrm{K}_{\mathcal{E}}$ in the following way: Take

$$\Phi(P) := \mathrm{argmin}\left\{D(Q\|\nu) : Q\in\mathcal{N}_P,\ \mathrm{supp}(Q)\cap\mathrm{supp}(P)=\emptyset\right\}.$$

Then $\Phi(P)$ is an element of $\overline{\mathcal{E}^{\mathcal{X}\setminus\mathrm{supp}(P)}}$, and if $P=u^+$ for some f.d.p. $u\in\partial\mathbf{U}_{\mathcal{N}}$, then $\Phi(P)=u^-$. More generally, if $P=v^+$ for some $v\in\partial\mathbf{U}_{\mathcal{N}}$, then $\Phi(v^+)$ equals the rI-projection of v^- onto $\mathcal{E}^{\mathcal{X}\setminus\mathrm{supp}(v^+)}$. This idea is useful to find the local maximizer of $\overline{D}_{\mathcal{E}}$ for low-dimensional exponential families, see Section 4.1.

3.5. The relation between $D(\cdot\|\mathcal{E})$ and $\overline{D}_{\mathcal{E}}$

The goal of this section is to generalize Theorem 3.16. The first step is to characterize the largest subsets of $\mathbf{P}(\mathcal{X})$ and $\partial\mathbf{U}_{\mathcal{N}}$ such that the two maps $\Psi_{\mathcal{E}}$ and $\Psi^+ : u\mapsto u^+$ induce mutually inverse bijections. On one side, the situation is clear: Lemma 3.7 says that the property $\Psi^+\circ\Psi_{\mathcal{E}}(P)=P$ characterizes projection points. The other side is given by the f.d.p.s, as Theorem 3.28 will show. The second step is to study how subsets of the sets of projection points and f.d.p.s behave under the maps $\Psi_{\mathcal{E}}$ and Ψ^+.

Theorem 3.28. *Let $\mathcal{E}$ be an exponential family with normal space $\mathcal{N}$. The maps $\Psi_\mathcal{E}$ and $\Psi^+ : u \mapsto u^+$ restrict to mutually inverse bijections between the set of proper projection points and the set of facial differences of projection points. $\Psi_\mathcal{E}$ maps the*

$$\text{set of } \begin{cases} \text{local maximizers} \\ \text{global maximizers} \\ \text{critical points} \\ \text{quasi-critical points} \end{cases} \text{ of } D_\mathcal{E} \text{ onto the set of } \begin{cases} \text{local maximizers} \\ \text{global maximizers} \\ \text{critical points} \\ \text{quasi-critical points} \end{cases} \text{ of } \overline{D}_\mathcal{E}.$$

For any projection point $P \in \mathbf{P}(\mathcal{X})$ and f.d.p. $u \in \partial\mathbf{U}_\mathcal{N}$ such that $P = u^+$,

$$D(P\|\mathcal{E}) = \log(1 + \exp(\overline{D}_\mathcal{E}(u))). \tag{3.14}$$

The proof of the statement about the local maximizers needs the following inequality, which is an analogue to Lemma 3.17:

Lemma 3.29. *$D_\mathcal{E}(P) \leq \log(1+\exp(\overline{D}_\mathcal{E}(\Psi_\mathcal{E}(P))))$ for all $P \in \mathbf{P}(\mathcal{X}) \setminus \overline{\mathcal{E}}$, with equality if and only if P is a projection point.*

Proof. Let $u = \Psi_\mathcal{E}(P)$, and let $\mathcal{E}'$ be the exponential family with reference measure ν and normal space $\mathcal{N}' = \mathbb{R}u$. Then $\mathcal{E} \subseteq \mathcal{E}'$, and u^+ is a projection point of $\mathcal{E}'$. Furthermore, $P_\mathcal{E} = P_{\mathcal{E}'}$, since $P - P_\mathcal{E} \in \mathcal{N}'$ and $P_\mathcal{E} \in \overline{\mathcal{E}'}$. Example 3.13 shows that $D(P\|\mathcal{E}') \leq D(u^+\|\mathcal{E}')$. Therefore,

$$\log(1 + \exp(\overline{D}_\mathcal{E}(\Psi_\mathcal{E}(P)))) = \log(1 + \exp(\overline{D}_{\mathcal{E}'}(u))) = D(u^+\|\mathcal{E}') \geq D(P\|\mathcal{E}') = D(P\|\mathcal{E}).$$

Equality is equivalent to $P = u^+$, and by Lemma 3.7 this means that P is a projection point. □

Proof of Theorem 3.28. The proof has four steps:

(1) *Projection points and f.d.p.s:* If P is a projection point, then $\Psi_\mathcal{E}(P)^-$ is also a projection point, and $\operatorname{supp}(\Psi_\mathcal{E}(P)) = \operatorname{supp}(P_\mathcal{E})$ is facial. By Lemma 3.25, projection points are mapped to f.d.p.s by $\Psi_\mathcal{E}$. Lemma 3.25 also shows that Ψ^+ maps f.d.p.s to projection points. By Lemma 3.7, $\Psi_\mathcal{E}$ is injective on the projection points, and Ψ^+ is surjective onto the projection points. Injectivity of Ψ^+ on the f.d.p.s follows from Lemma 3.25: Let $u, v \in \partial\mathbf{U}_\mathcal{N}$ be f.d.p.s such that $u^+ = v^+$, and let $P_\mathcal{E}$ be the rI-projection of u^+ onto $\mathcal{E}$. Then $P_\mathcal{E}$ is a convex combination of u^+ and u^- as well as a convex combination of u^+ and v^-. Since u^- and v^- are probability measures on $\mathcal{X} \setminus \operatorname{supp}(u^+)$, they must be equal. This proves the statement about the projection points and f.d.p.s. Equation (3.14) is a consequence of Lemma 3.17 or Lemma 3.29.

(2) *The quasi-critical points:* By Lemma 3.24, a projection point P is a quasi-critical point of $D_\mathcal{E}$ if and only if the f.d.p. $\Psi_\mathcal{E}(P)$ is a quasi-critical point of $\overline{D}_\mathcal{E}$.

(3) *The critical points:* Let $P \in \mathbf{P}(\mathcal{X})$ be a quasi-critical point of $D_\mathcal{E}$ with rI-projection $P_\mathcal{E}$. From Theorem 3.16 applied to the exponential family $\mathcal{E}^{P_\mathcal{E}}$ and equation (3.14) it follows that P satisfies inequality (3.2) if and only if $u = \Psi_\mathcal{E}(P)$ satisfies (3.9).

(4) *The maximizers:* The statement about the global maximizers is Theorem 3.16. Let $V \subseteq \partial\mathbf{U}_\mathcal{N}$ be a neighbourhood of u such that $\overline{D}_\mathcal{E}(v) \leq \overline{D}_\mathcal{E}(u)$ for all $v \in V$, and

let $U = \Psi_{\mathcal{E}}^{-1}(V)$. Since u is a critical point of $\overline{D}_{\mathcal{E}}$, the open set U contains u^+. By Lemmas 3.17 and 3.29, if $Q \in U$, then

$$D(Q\|\mathcal{E}) \le \log(1+\exp(\overline{D}_{\mathcal{E}}(\Psi_{\mathcal{E}}(Q)))) \le \log(1+\exp(\overline{D}_{\mathcal{E}}(u))) \le D(u^+\|\mathcal{E}).$$

Conversely, if $P \in \mathbf{P}(\mathcal{X})$ is a local maximizer of $D_{\mathcal{E}}$, then let $U \subseteq \mathbf{P}(\mathcal{X})$ be a neighbourhood of P such that $D(Q\|\mathcal{E}) \le D(P\|\mathcal{E})$ for all $Q \in U$. The open set $V = (\Psi^+)^{-1}(Q)$ is a neighbourhood of $\Psi_{\mathcal{E}}(P)$. By Lemmas 3.17 and 3.14, if $u \in V$, then

$$\overline{D}_{\mathcal{E}}(u) \le \log(\exp(D(u^+\|\mathcal{E})) - 1) \le \log(\exp(D(P\|\mathcal{E})) - 1) = \overline{D}_{\mathcal{E}}(\Psi_{\mathcal{E}}(u)). \qquad \square$$

Theorem 3.28 makes it possible to relate properties of the local maximizers of $\overline{D}_{\mathcal{E}}$ and $D_{\mathcal{E}}$. For example, the following lemma is a consequence of Lemma 3.5. It is instructive to prove it directly, though. The proof should also be compared to the proof of Lemma 3.26.

Lemma 3.30. *Let $\mathcal{E}$ be an exponential family with sufficient statistics A, and let u be a local maximizer of $\overline{D}_{\mathcal{E}}$. Then the set $\{A_x : u(x) > 0\}$ is affinely independent. In particular, the support of u^+ has cardinality at most* $\dim \mathcal{E} + 1$.

Proof. Let $\mathcal{Z} = \operatorname{supp}(u^+)$. If u is a local maximizer of $\overline{D}_{\mathcal{E}}$, then u^+ is a local maximizer of the strictly convex function $D(\cdot\|\nu)$ restricted to the polytope $\mathbf{P}(\mathcal{Z}) \cap \mathcal{N}_{u^-}$. Therefore, u^+ is a vertex of this polytope, so $\{A_x : u(x) > 0\}$ is affinely independent. $\square$

Remark 3.31. It is an interesting question when the rI-projection $P_{\mathcal{E}}$ of a local maximizer P lies in $\mathcal{E}$. More generally one can ask for the support of $P_{\mathcal{E}}$. Since $\operatorname{supp}(P_{\mathcal{E}}) = \operatorname{supp}(\Psi_{\mathcal{E}}(P))$ it is equivalent to ask for the support of a local maximizer of $\overline{D}_{\mathcal{E}}$ by Theorem 3.28. In many cases the support of $P_{\mathcal{E}}$ is equal to $\mathcal{X}$, but the construction of Example 3.13 shows that $P_{\mathcal{E}}$ can have any support (of cardinality at least two).

3.6. Computing the critical points

The most direct approach to finding the local maximizers of $\overline{D}_{\mathcal{E}}$ is to solve the critical equations (3.8). The function $\overline{D}_{\mathcal{E}}$ is not smooth, since the function $p \mapsto p\log|p|$ is not smooth at $p = 0$. This makes it difficult to treat the first order conditions of $\overline{D}_{\mathcal{E}}$. However, $\overline{D}_{\mathcal{E}}$ is smooth when the sign vector is restricted: For any $v \in \mathcal{N}$ let $\operatorname{sgn}(v) = (\operatorname{sgn}(v_x))_{x\in\mathcal{X}} \in \{0, \pm 1\}^{\mathcal{X}}$ be the sign vector of v. Then the restriction of $\overline{D}_{\mathcal{E}}$ to $\mathcal{N}^{\operatorname{sgn}(u)} = \{v \in \mathcal{N} : \operatorname{sgn}(v) = \operatorname{sgn}(u)\}$ is a smooth function for any $u \in \mathcal{N}$. Proposition 3.21 (i) and Lemma 3.30 can also be seen as conditions on the sign vector $\operatorname{sgn}(u)$ of u:

Definition 3.32. Let $\mathcal{N} \subseteq \mathbb{R}^{\mathcal{X}}$. A sign vector $\sigma \in \operatorname{sgn}(\mathcal{N})$ is *facial* if its support is facial. σ is a *quasi-critical sign vector* of $\mathcal{N}$ if it satisfies

$$v(\sigma = 0) = 0, \qquad \text{for all } v \in \mathcal{N}.$$

If σ is quasi-critical and if $\{A_x : \sigma_x > 0\}$ is affinely independent, then σ is a *critical sign vector*. By Lemma 3.24, quasi-critical and critical sign vectors are facial.

By what was said until now, the following strategy can be used to compute the local maximizers of $\overline{D}_{\mathcal{E}}$:

1. Compute all critical sign vectors.
2. For any critical sign vector $\sigma = (\sigma_x)_{x\in\mathcal{X}} \in \operatorname{sgn}(\mathcal{N})$ find all $u \in \mathcal{N}^\sigma$ that solve the equations
$$\sum_{x:\sigma_x\neq 0} v(x) \log \frac{\sigma_x u(x)}{\nu_x} = 0 \tag{3.15}$$
for all $v \in \mathcal{N}$ satisfying $\operatorname{supp}(v) \subseteq \operatorname{supp}(u)$ and $d'_u(v) = v(u > 0) + v^+(u = 0) = 0$.
3. Determine which quasi-critical points are critical.
4. Check which critical points are local maximizers.

If only the global maximum is of interest, then it is sufficient to compute the value of $\overline{D}_{\mathcal{E}}$ at each quasi-critical point. Similarly, the f.d.p.s can be computed by considering all facial sign vectors. The set of all sign vectors occurring in a vector space (in this case $\mathcal{N}$) forms a (realizable) oriented matroid. The basic facts about sign vectors and oriented matroids which are needed in the following are collected in Appendix A.2.

The number of sign vectors of $\mathcal{N}$ can be very large. A naive upper bound is 3^N, where $N = |\mathcal{X}|$. In contrast, the number of subsets of $\mathcal{X}$ is 2^N. Therefore, one might think that there are much more quasi-critical sign vectors than possible support sets for local maximizers of $D_{\mathcal{E}}$. The following lemma states that this is not true:

Lemma 3.33. *Let $\sigma, \tau \in \operatorname{sgn}(\mathcal{N})$ be two facial sign vectors. If $\sigma^+ = \tau^+$, then $\sigma = \tau$.*

Proof. By Lemma 2.30, if $\operatorname{supp}(\sigma^+) \subseteq \operatorname{supp}(\tau)$, then $\operatorname{supp}(\sigma^-) \subseteq \operatorname{supp}(\tau)$, and vice versa. □

Let σ be the sign vector of some $u_0 \in \partial\mathbf{U}_{\mathcal{N}}$. Write $\mathcal{Y} := \operatorname{supp}(\sigma) = \operatorname{supp}(u_0)$. Define $d^\sigma(v) := \sum_{x:\sigma_x>0} v(x)$. This implies $d^\sigma(v) = d'_{u_0}(v)$ whenever $\operatorname{supp}(v) \subseteq \operatorname{supp}(u_0) = \operatorname{supp}(\sigma)$. Let

$$K^\sigma := \big\{v \in \mathcal{N} : d^\sigma(v) = 0 \text{ and } \operatorname{supp}(v) \subseteq \operatorname{supp}(\sigma)\big\}.$$

If $u \in \partial\mathbf{U}_{\mathcal{N}}$ satisfies $\operatorname{sgn}(u) = \sigma$, then $u - u_0 \in K^\sigma$. Exponentiating the critical equations (3.15) proves the following result:

Proposition 3.34. *Let σ be the sign vector of $u \in \partial\mathbf{U}_{\mathcal{N}}$, and suppose that u satisfies*

$$\prod_{x\in\mathcal{Y}} \left(\frac{\sigma_x u(x)}{\nu_x}\right)^{v^+(x)} = \prod_{x\in\mathcal{Y}} \left(\frac{\sigma_x u(x)}{\nu_x}\right)^{v^-(x)}, \quad \text{for all } v \in K^\sigma. \tag{3.16}$$

If σ is facial, then u is a f.d.p. If σ is quasi-critical, then u is a quasi-critical point of $\overline{D}_{\mathcal{E}}$. Conversely, if u is a f.d.p., then u satisfies (3.16).

Remark 3.35. The system of equations (3.16) contains infinitely many equations. The derivation of (3.16) from (3.15) (which is linear in v) shows that it is enough to consider these equations for v from a spanning set of K^σ. This is due to the fact that these equations correspond to the directional derivatives of a smooth function. Hence a finite number of equations is sufficient.

If $v \in K^\sigma$ happens to be an integer vector, then (3.16) is a polynomial equation. Therefore, if

$$K^\sigma_\mathbb{Z} := K^\sigma \cap \mathbb{Z}^\mathcal{X}$$

contains a spanning set of K^σ, then the critical points with sign vector σ are solutions to polynomial equations.

Lemma 3.36. *If $\mathcal{E}$ is an algebraic exponential family, then $K^\sigma_\mathbb{Z}$ spans K^σ for all sign vectors $\sigma \in \operatorname{sgn}(\mathcal{N})$. Conversely, if the codimension of $\mathcal{E}$ is at least two and if $K^\sigma_\mathbb{Z}$ spans K^σ for all sign vectors $\sigma \in \operatorname{sgn}(\mathcal{N})$, then $\mathcal{E}$ is algebraic.*

Proof. Assume that $\mathcal{N}$ has an integer basis. For any σ the vector space K^σ is the solution set of a system of linear equations with integer coefficients; therefore, K^σ also has an integer basis.

For the other direction, if $\dim(\mathcal{N}) \geq 2$, then there exist two sign vectors σ, σ' with maximal support such that $\sigma \neq \sigma' \neq -\sigma$. Then K^σ and $K^{\sigma'}$ are two different vector subspaces of codimension one of $\mathcal{N}$. If K^σ and $K^{\sigma'}$ both have an integer basis, then $\mathcal{N} = K^\sigma + K^{\sigma'}$ also has an integer basis. □

Proposition 3.37. *Let $\mathcal{E}$ be an algebraic exponential family. Let σ be the sign vector of $u \in \partial\mathbf{U}_\mathcal{N}$, and suppose that u satisfies the polynomial equations*

$$u^{v^+}\nu^{v^-} = u^{v^-}\nu^{v^+}, \quad \text{for all } v \in K^\sigma_\mathbb{Z}, \tag{3.17}$$

If σ is facial, then u is a f.d.p. If σ is quasi-critical, then u is a quasi-critical point of $\overline{D}_\mathcal{E}$. Conversely, if u is a f.d.p., then u satisfies (3.17).

Proof. It only remains to see that the factors $\sigma_x^{v^\pm(x)}$ cancel each other; then the statements follow from Proposition 3.34, Lemma 3.36 and the discussion above. If $v \in K^\sigma_\mathbb{Z}$, then $v(\sigma < 0) = v(\sigma \neq 0) - v(\sigma > 0) = 0$. Hence $v^+(\sigma < 0) + v^-(\sigma < 0) = 0$, and so

$$\prod_{x:v(x)>0} (\sigma_x)^{v(x)} = (-1)^{v^+(\sigma<0)} = (-1)^{v^-(\sigma<0)} = \prod_{x:v(x)<0} (\sigma_x)^{-v(x)}. \qquad \square$$

In the rest of this section only algebraic exponential families will be considered. The study of polynomials is in many respects easier when the field is algebraically closed (but not always, see Remark 3.39). Therefore, it is convenient to interpret equations (3.17) as complex equations in the variables $u(x)$. Of course, only real solutions with the right sign pattern are candidate solutions of the original maximization problem.

Fix a critical sign vector σ, and let A be a sufficient statistics of $\mathcal{E}$ satisfying the statements of Lemma 2.9. Let I_2^σ be the ideal generated by all equations (3.17) in

the polynomial ring $\mathbb{C}[u(x) : x \in \mathcal{Y}]$ with one variable for each $x \in \mathcal{Y}$. Similarly, let $I_1^\sigma \subseteq \mathbb{C}[u(x) : x \in \mathcal{Y}]$ be the ideal generated by the equations

$$\sum_{x \in \mathcal{Y}} A_{i,x} u(x) = 0, \qquad \text{for all } i.$$

Finally let $I^\sigma := I_1^\sigma + I_2^\sigma$. The set of all common complex solutions of all equations in I^σ is an algebraic subvariety of $\mathbb{C}^\mathcal{Y}$ and is denoted by X^σ.

Remark 3.38. It is possible to add the equation $d_\sigma - 1 = 0$ to any of the ideals defined above. Without this equation the above ideals are all homogeneous and define projective varieties. As in Remark 2.33 the equation $d_u - 1 = 0$ corresponds to a normalization condition and can be ignored for theoretical purposes, but it is useful in applications to eliminate one of the variables.

Both ideals I_1^σ and I_2^σ taken for themselves are easy to solve: I_1^σ corresponds to a system of linear equations, so it can be treated by the methods of linear algebra. On the other hand, I_2^σ is a binomial ideal, so the discussion of Section 2.3 applies. The relation between the binomial ideals $I_\nu(B)$ of Section 2.3 and the binomial ideals I_2^σ will become clearer in the next section.

The sum of a linear ideal and a binomial ideal can be arbitrarily complicated. In fact, it is easy to see that any system of polynomial equations can be reparametrized as a combination of linear and binomial equations: For example, a polynomial equation $\sum_i m_i = 0$ with arbitrary monomials m_i, is equivalent to the system of equations

$$\begin{aligned} z_i - m_i &= 0, \text{ for all } i, \\ \sum_i z_i &= 0, \end{aligned}$$

where one additional variable z_i has been introduced for every monomial. Still, the two ideals I_1^σ and I_2^σ under consideration here are closely related, and one may try to exploit this relationship.

X^σ equals the intersection of X_1^σ and X_2^σ, where X_1^σ and X_2^σ are the varieties of I_1^σ and I_2^σ, respectively. The variety X_1^σ is easy to determine: It is given by the (complex) kernel of A restricted to $\mathcal{Y}$:

$$X_1^\sigma = \ker_\mathbb{C} A \cap \mathbb{C}^\mathcal{Y} = \ker_\mathbb{C} A_\mathcal{Y},$$

where $A_\mathcal{Y}$ is the submatrix of A with columns A_x for $x \in \mathcal{Y}$. The variety X_2^σ is slightly more complicated. Theorem 2.35 implies that I_2^σ is a prime ideal, so X_2^σ is a toric variety. In particular X_2^σ is irreducible.

For any subset $B \subseteq K_\mathbb{Z}^\sigma$ let $I_2^\sigma(B)$ be the ideal generated by the polynomials

$$\prod_{x \in \mathcal{Y}} \left(\frac{u(x)}{\nu_x} \right)^{v^+(x)} - \prod_{x \in \mathcal{Y}} \left(\frac{u(x)}{\nu_x} \right)^{v^-(x)}$$

for $v \in B$. If B generates $K_\mathbb{Z}^\sigma$ (as an abelian group), then Proposition 2.42 says that any $u \in \partial \mathbf{U}_\mathcal{N}$ with sign vector σ solves $I_2^\sigma(B)$ if and only if u solves I_2^σ. Therefore, one may

replace I_2^σ with $I_2^\sigma(B)$ in the following calculations. Even more is true: Remark 3.35 applies, hence it suffices if B is a basis of K^σ. However, for computational reasons it is preferable if B is a circuit basis or even a Markov basis, as discussed in Section 2.3.

Although X_1^σ and X_2^σ are both irreducible, in general X^σ may be reducible. This means that X^σ is a finite union of irreducible components $X^\sigma = V_1^\sigma \cup \cdots \cup V_c^\sigma$. To each of these components V_i^σ corresponds a polynomial ideal I_i^σ, and $u \in X^\sigma$ if and only if u solves (at least) one of these ideals, i.e. $I^\sigma = I_1^\sigma \cap \cdots \cap I_c^\sigma$. The procedure to obtain the ideals I_i^σ is called *primary decomposition*. In fact, a primary decomposition of an ideal can give much more information than an irreducible decomposition of varieties, since an ideal may contain more information than a variety, in the sense that different ideals can define the same variety. These fine points play no role in the following, since, in the end, only the varieties are of interest here.

There are symbolic algorithms [33] as well as numerical algorithms [65] for primary decomposition. The symbolic algorithms are implemented, for example, in the computer algebra systems SINGULAR [34] or Macaulay2 [32]. An implementation of the numerical algorithms is Bertini [11]. The applicability of numerical algorithms is discussed below in Remark 3.41.

Remark 3.39. Symbolic and numerical algorithms differ in their requirements on the ground field: Symbolic algorithms cannot work with arbitrary real or complex numbers, since these cannot be represented exactly in a computer. The exact representation of algebraic numbers over $\mathbb{Q}$ is possible, but many implementations of the symbolic algorithms contend themselves to compute a primary decomposition over $\mathbb{Q}$. It is then possible to deduce the full primary decomposition by adjoining certain algebraic numbers to $\mathbb{Q}$ (and this last step is usually not difficult). On the other hand, the numerical algorithms usually work with complex numbers, or with floating point approximations, to be precise. Most algorithms rely on homotopy continuation of solutions of polynomial equations, and hence they need a ground field that is algebraically closed as well as topologically complete.

Primary decompositions can be computationally challenging, even for moderately sized systems of polynomial equations. Therefore, it is important to choose the set of equations wisely[1] before invoking a computer algebra system. The choice of the binomials discussed above is one aspect. Another possibility is to incorporate further knowledge about the sign vector in the algebraic equations. The construction of the ideal I^σ already ensures that the support of any point in X^σ is contained in $\mathcal{Y}$. While it is difficult to find only real solutions satisfying the right sign conditions by purely algebraic means, there is an algebraic method to discard some solutions with a too small support.

Let $x \in \mathcal{Y}$. For every irreducible component V_i^σ of X^σ there are two alternatives:

- Either $u(x) = 0$ for all $u \in V_i^\sigma$. In this case, $\operatorname{sgn}(u) \neq \sigma$ for all $u \in V_i^\sigma$,
- or $u(x) = 0$ holds only on a subset of V_i^σ of measure zero.

[1] or use trial and error

The reason for this is that the equation $u(x) = 0$ defines a closed subset of V_i^σ, and either this closed subset is all of V_i^σ, or it has codimension one (this argument needs the irreducibility of V_i^σ).

When computing the primary decomposition the irreducible components of the first kind can be excluded algebraically by saturation (see Section 2.3): Namely, the variety corresponding to the saturated ideal

$$I^\sigma : \Big(\prod_{x\in\mathcal{Y}} u(x)\Big)^\infty$$

consists only of those irreducible components of X^σ that are not contained in any coordinate hyperplane. Similarly, one may saturate I^σ by the polynomial $d^\sigma(u)$, since any solution u with $\operatorname{sgn}(u) = \sigma$ has $0 \neq d(u) = d^\sigma(u)$.

Both SINGULAR and Macaulay2 provide routines to compute saturations. Still, saturations can be a computationally difficult. If different saturations have to be performed, then changing the order of these saturations can decisively speed up the calculation. For example, instead of saturating with respect to all monomials at the same time, it may be advantageous to saturate with respect to the variables $u(x)$, $x \in \mathcal{Y}$, one by one.

If the saturation can be computed, then it may reduce the complexity of subsequent symbolic calculations. Furthermore, saturation removes irreducible components from the solution set that do not contain a point with the right sign vector, thus simplifying the analysis of the results.

Assume that the irreducible decomposition $X^\sigma = V_1^\sigma \cup \cdots \cup V_c^\sigma$ can be found. If an irreducible component V_i^σ is zero-dimensional, then it consists of only one point (at least over an algebraically closed field like the complex numbers, cf. Remark 3.39), and it is easy to check whether this unique element $u \in V_i^\sigma$ satisfies $\operatorname{sgn}(u) = \sigma$. Components of positive dimension may arise, however. In this case it is difficult to decide whether these components contain elements u satisfying $\operatorname{sgn}(u) = \sigma$. Fortunately, in many cases this information is not required if only the global maximum is of interest:

Theorem 3.40. *Let u be an element of an irreducible component V of X^σ such that $d_\sigma(u) = 1$. Suppose there exists a real $u_0 \in V$ such that $d_\sigma(u_0) = 1$ and $\operatorname{sgn}(u_0) = \sigma$. Then*

$$\overline{D}_{\mathcal{E}}(u_0) = \sum_{x\in\mathcal{Y}} \Re(u(x)) \log \frac{|u(x)|}{\nu_x}.$$

Proof. Let

$$V' := \{v \in V : v(x) \neq 0 \text{ for all } x \in \mathcal{Y}, \text{ and } d_\sigma(v) \neq 0\}.$$

Then V' is a Zariski-open subset of V, hence V' is irreducible. This implies that V' is pathwise connected, so there exists a smooth path $\gamma : [0, 1] \to V'$ from u to u_0. This is obvious if V' is regular, since V' is a locally pathwise connected and connected complex manifold in this case. In the general case, all regular points can be connected by a smooth path. Finally, every singular point p can be linked by a smooth path

to some regular point in any neighbourhood of p. This path can be chosen such that $d_\sigma(\gamma_t) = 1$ for all $t \in [0,1]$, cf. Remark 3.38.

Let $u \in V'$, and fix a branch of the complex logarithm. For every $x \in \mathcal{Y}$ the logarithm can be continued analytically along the path $t \mapsto \gamma_t(x)$ to a map $t \mapsto \log^{t,x}(\gamma_t(x))$, such that $\log^{0,x}(\sigma_x\gamma_0(x)) = \log(\sigma_x u(x))$. Let $K^\sigma_{\mathbb{C}}$ be the complexification of K^σ. For every $t \in [0,1]$ define a linear functional $s_t : K^\sigma_{\mathbb{C}} \to \mathbb{C}$ via

$$s_t(v) = \frac{1}{2\pi\mathfrak{i}} \sum_{x\in\mathcal{Y}} v(x) \log^{t,x} \frac{\sigma_x\gamma_t(x)}{\nu_x}.$$

The definition of X^σ shows that $\exp(2\pi\mathfrak{i}s_t(v)) = 1$ for all $t \in [0,1]$ and all $v \in K^\sigma_{\mathbb{Z}}$. Hence s_t takes only integer values on $K^\sigma_{\mathbb{Z}}$ and can be identified with an element of the dual lattice $K^{\sigma*}_{\mathbb{Z}}$ of $K^\sigma_{\mathbb{Z}}$. Since $K^{\sigma*}_{\mathbb{Z}}$ is a discrete subset of the dual vector space $K^{\sigma*}_{\mathbb{C}}$ and since the map $t \mapsto s_t$ is continuous, s_t is constant along γ.

Consider the function

$$f(t) = \sum_{x\in\mathcal{Y}} \gamma_t(x) \log^{t,x} \left(\frac{\sigma_x\gamma_t(x)}{\nu_x} \right).$$

Its derivative is

$$f'(t) = \sum_{x\in\mathcal{Y}} \gamma_t'(x) \log^{t,x} \left(\frac{\sigma_x\gamma_t(x)}{\nu_x} \right) = 2\pi\mathfrak{i}s_0(\gamma_t'),$$

where $\gamma_t'(x) = \frac{\partial}{\partial t}\gamma_t(x) \in K^\sigma_{\mathbb{C}}$. Hence, $f(1) - f(0) = 2\pi\mathfrak{i}s_0(\gamma_1 - \gamma_0)$. In other words,

$$\sum_{x\in\mathcal{Y}} u_0(x) \log^{1,x} \frac{\sigma_x u_0(x)}{\nu_x} = \sum_{x\in\mathcal{Y}} u(x) \log^{0,x} \frac{\sigma_x u(x)}{\nu_x} + 2\pi\mathfrak{i}s_0(u_0 - u).$$

If $\log^{1,x}(\sigma_x u_0(x)) = \log(\sigma_x u_0(x)) + 2\pi i k_x$ with $k_x \in \mathbb{Z}$, then

$$\overline{D}_{\mathcal{E}}(u_0) = f(0) + 2\pi\mathfrak{i} \left(s_0(u_0 - u) - \sum_{x\in\mathcal{Y}} u_0(x)k_x \right).$$

Taking the real parts of this equation gives

$$\begin{aligned}\overline{D}_{\mathcal{E}}(u_0) &= \Re(f(0)) + 2\pi s_0(\Im(u)) = \Re(f(0)) - \mathfrak{i}\sum_{x\in\mathcal{Y}} \Im(u(x)) \log \frac{\sigma_x u(x)}{\nu_x} \\ &= \Re\left(\sum_{x\in\mathcal{Y}} \Re(u(x)) \log \frac{\sigma_x u(x)}{\nu_x} \right) = \sum_{x\in\mathcal{Y}} \Re(u(x)) \log \frac{|u(x)|}{\nu_x}.\end{aligned}$$

By continuity this formula continues to hold when u belongs to the closure of V', which equals V. □

If only the global maximum of $\overline{D}_{\mathcal{E}}$ is of interest, then the theorem implies that for many irreducible components of some X^σ it is enough to know one point u. Only if $\sum_{x\in\mathcal{Y}} \Re(u(x)) \log \frac{|u(x)|}{\nu_x}$ is exceptionally large, then it is necessary to analyze this irreducible component further and see if there is a real point u_0 from the same irreducible component that satisfies the sign condition.

Remark 3.41. The above theorem makes it possible to use methods of numerical algebraic geometry [65], like those implemented in Bertini [11]. These methods can determine the number of irreducible components and their dimensions. In addition, it is possible to sample points from any irreducible component. In fact, each component is represented by a so-called *witness set*, a set of elements of this component. These points can be used to numerically evaluate $\overline{D}_{\mathcal{E}}$.

Example 3.42. The above ideas can be applied to the pair interaction model of four binary random variables. The maximization problem of this model is related to orthogonal Latin squares: A global maximizer of the information divergence from the homogeneous pair interaction model with variables of size $N_1 = k$ is easy to find if two orthogonal Latin squares of size k exist, and in this case the maximum value of $D_{\mathcal{E}}$ equals $2\log(k)$, see [51]. Otherwise the maximum value of $D_{\mathcal{E}}$ is strictly less than $2\log(k)$. From this point of view, the following computation is a very complicated proof of the trivial fact that there are no two orthogonal Latin squares of size two.

A sufficient statistics and a basis of the normal space $\mathcal{N}$ were given in Section 2.4. The software package `TOPCOM` is used to calculate the oriented circuits of $\mathcal{N}$, from which all sign vectors are computed by composition. Up to symmetry there are 73 different sign vectors occurring in $\mathcal{N}$. Here, the symmetry group of the model is generated by the permutations of the four binary units and the relabellings $1 \leftrightarrow 2$ of each unit.

From these 73 sign vectors only 20 are critical. The sign vectors of small support are easy to handle: There are two critical sign vectors σ_1, σ_2 with support of cardinality eight. They are oriented circuits, which implies that there are two unique elements $u_1, u_2 \in \partial\mathbf{U}_{\mathcal{N}}$ such that $\mathrm{sgn}(u_i) = \sigma_i$, $i = 1, 2$. They satisfy $\overline{D}_{\mathcal{E}}(u_i) = 0$, so they are surely not global maximizers. There are three critical sign vectors with support of cardinality twelve. Let σ be one of these. Then the restriction $\mathrm{supp}(u) \subseteq \mathrm{supp}(\sigma)$ selects a two-dimensional subspace of $\mathcal{N}$, and it is easy to see that $\overline{D}_{\mathcal{E}} = 0$ on this subspace. The reason for this is that there is a permutation τ of $\mathrm{supp}(\sigma)$ such that $v(\tau(x)) = -v(x)$ for all $x \in \mathrm{supp}(\sigma)$ and all $v \in \mathcal{N}$ that satisfy $\mathrm{supp}(v) \subseteq \mathrm{supp}(\sigma)$.

There remain 15 critical sign vectors with full support. For every such sign vector σ, the system of the algebraic equations in $I^\sigma = I_1^\sigma + I_2^\sigma$ has to be solved. To reduce the number of equations and the number of variables one may parametrize the solution set X_1^σ of I_1^σ by finding a basis $u_1, \ldots, u_5$ of $\mathcal{N}$. Then the parametrization $u(\lambda_1, \ldots, \lambda_5) = \sum_{i=1}^5 \lambda_i u_i$ of X_1^σ is plugged into the equations of I_2^σ. This yields an ideal in $\mathbb{C}[\lambda_1, \ldots, \lambda_5]$. Some of these ideals are at the limit of what today's desktop computers can handle. Therefore, care has to be taken how to formulate these equations. The general strategy is the following:

1. Compute a basis $v_1, \ldots, v_4$ of $K_{\mathbb{Z}}^\sigma$ using a Gram-Schmidt-like algorithm: Renum-

ber the u_i such that $d_\sigma(u_5) \neq 0$ and let

$$v_i := \frac{d_\sigma(u_5)}{g} u_i - \frac{d_\sigma(u_i)}{g} u_5,$$

where $g = \gcd(d_\sigma(u_5), d_\sigma(u_i))$.

2. Let I be the ideal in the variables $\lambda_1, \ldots, \lambda_5$ generated by the equations

$$\prod_{x:v_i>0} u(x)^{v_i(x)} - \prod_{x:v_i<0} u(x)^{-v_i(x)}, \text{ for } i = 1, \ldots, 4,$$

where $u(x) = \sum_{i=1}^5 \lambda_i u_i(x)$, and compute the saturation $J = I : \left(\prod_{x\in\mathcal{X}} u(x)\right)^\infty$.

3. Compute a primary decomposition of J.

The ideal I in the second step corresponds to the ideal $I_2^\sigma(B)$ defined above for the basis $B = \{v_1, \ldots, v_4\}$, where the variables $u(x)$ have been restricted to the linear subspace $\ker_{\mathbb{C}} A$. As discussed in Section 2.3 it would be better to replace B with a circuit basis or a Markov basis of $K_{\mathbb{Z}}^\sigma$. However, a basis proved to be sufficient. To automatize the algorithm a C++ program was written that takes a sign vector and computes the generators of I in Singular syntax. Using a circuit basis or a Markov basis would have made it necessary to use external programs or libraries, and it was easier to implement the above Gram-Schmidt-like algorithm.

Unfortunately, this simple algorithm does not work for all sign vectors. Some further tricks are needed to compute the primary decomposition within a reasonable time: The basis of adjacent minors of $\mathcal{N}$ (see Theorem 2.48) is given by the rows $u_1, \ldots, u_5$ of the matrix

$$\begin{pmatrix}
1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\
1 & 0 & -1 & 0 & -1 & 0 & 1 & 0 & -1 & 0 & 1 & 0 & -1 & 0 \\
1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & -1 & 1 & 0 & 0 & 1 & -1 & 0 & 0 \\
1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & -1 & 1 & 1 & -1 & 0 & 0 & 0 & 0 \\
1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}.$$

This basis has the following property: Let $u = \sum_{i=1}^5 \lambda_i u_i$. If $\lambda_j = 0$ for some $j = 2, 3, 4, 5$, then $\overline{D}_{\mathcal{E}}(u) = 0$, because in this case there is a bijection between the positive and negative entries of u such that corresponding entries have the same absolute value. Hence, in order to determine the global maximizers of this model, one may saturate J by the product $\lambda_2\lambda_3\lambda_4\lambda_5$.

Replacing J by $(J : (\lambda_2\lambda_3\lambda_4\lambda_5)^\infty)$ makes it possible to solve all but one system of equations. For the last sign vector σ a special treatment is necessary: The complexity of the above algorithm depends on the chosen basis v_1, v_2, v_3, v_4 of $K_{\mathbb{Z}}^\sigma$. The ℓ_1-norm of each vector v_i equals twice the degree of the corresponding equation. Thus it is advisable to choose the vectors v_1, v_2, v_3, v_4 as short as possible. As a first approximation, one may try to use a basis of prime circuit vectors. This approach provides

a basis v_1, v_2, v_3, v_4 for $K_{\mathbb{Z}}^{\sigma}$ of the last sign vector, such that the rest of the algorithm sketched above works (an alternative would be to replace $\{v_1, v_2, v_3, v_4\}$ with a prime circuit basis or a Markov basis of $K_{\mathbb{Z}}^{\sigma}$, cf. Section 2.3).

The calculations were performed with the help of SINGULAR. The primary decompositions were done using the algorithm of Gianni, Trager and Zacharias (GTZ) implemented in the SINGULAR library `primdec.lib` [56]. The following theorem sums up the results, which confirm a conjecture by Thomas Kahle (personal communication):

Theorem 3.43. *The pair interaction model with four binary units has, up to symmetry, a unique maximizer of the information divergence, which is the uniform distribution over the states* 1112*,* 1121*,* 1211*,* 2111 *and* 2222*. The maximal value of* $\overline{D}_{\mathcal{E}}$ *is* $\log 3 - \frac{1}{3}\log 5 \approx 0.56$*, and it is attained at*

$$u = \frac{1}{15}(-5,\ 3,\ 3, -1,\ 3, -1, -1, -1, 3, -1, -1, -1, -1, -1, -1,\ 3).$$

The maximum value of $D_{\mathcal{E}}$ *is* $\log(1 + 3 \cdot 5^{-\frac{1}{3}}) \approx 1.01$.

3.7. Computing the projection points

The results of Section 3.2 motivate a second method for computing the maximizers of $D_{\mathcal{E}}$. This method is actually more elementary than solving the critical equations of $\overline{D}_{\mathcal{E}}$ as in the previous section. Knowing the first order conditions of $\overline{D}_{\mathcal{E}}$ gives additional insight into the method presented in this section.

Let P be a projection point of the exponential family $\mathcal{E}$, let $u = \Psi_{\mathcal{E}}(P)$, and let $P_{\mathcal{E}}$ be the rI-projection of P and u^-. Then

$$u(x) = \begin{cases} \frac{1}{\mu} P_{\mathcal{E}}(x), & \text{if } u(x) > 0, \\ -\frac{1}{1-\mu} P_{\mathcal{E}}(x), & \text{if } u(x) < 0, \end{cases} \tag{3.18}$$

where $\mu = P_{\mathcal{E}}(\operatorname{supp}(P))$. Let $\mathcal{Y} = \operatorname{supp}(P_{\mathcal{E}})$, and let A be a sufficient statistics of $\mathcal{E}$. Theorem 2.29 implies that there exist $\xi_1, \ldots, \xi_h > 0$ such that

$$P_{\mathcal{E}}(x) = c' \nu_x \prod_{i=1}^{h} \xi_i^{A_{i,x}}, \qquad \text{for all } x \in \mathcal{Y}. \tag{3.19}$$

Here $c' > 0$ is a normalization constant.

Define an $(h+1) \times \mathcal{Y}$-matrix A^{σ} as follows: Take the columns A_x of A for $x \in \mathcal{Y}$, then add a zeroth row with entries

$$A^{\sigma}_{0,x} := \frac{1}{2}(1 - \sigma_x) \in \{0, 1\}, \qquad \text{for } x \in \mathcal{Y}.$$

Then (3.18) and (3.19) together show that u has the form

$$u(x) = c\nu_x \prod_{i=0}^{h} \xi_i^{A^\sigma_{i,x}}, \quad \text{for all } x \in \mathcal{Y}. \tag{3.20}$$

Here, $c = \frac{c'}{\mu}$, $\xi_0 = -\frac{\mu}{1-\mu} < 0$, and all the other parameters are positive. If A satisfies one of the conditions of Lemma 2.9, then c may be assumed to be one and omitted. The projection points of $\mathcal{E}$ can be found by plugging the parametrization (3.20) into the equation $Au = 0$ and solving for the ξ_i.

Again, this method simplifies if $\mathcal{E}$ is algebraic. In this case it is possible to choose a sufficient statistics $A \in \mathbb{N}_0^{h \times \mathcal{X}}$ which has only nonnegative integer entries. This nonnegativity requirement can be achieved by adding a suitable multiple of $\mathbf{1}$ to each row of A. With such a choice the parametrization (3.20) is monomial (in the algebraic sense), so the equation $Au = 0$ is equivalent to h polynomial equations in the $h + 1$ parameters $\xi_0, \ldots, \xi_h$.

This method is linked to the ideal I_2^σ of the previous section. As stated there, I_2^σ is a toric ideal, hence it defines a toric variety. Every toric variety has a monomial parametrization, which induces the monomial parametrization (3.20). Unfortunately, in general this monomial parametrization is not surjective, see [44]. The argument leading to equation (3.20) shows that it is "surjective enough," though.

The polynomial equations obtained from $Au = 0$ and the monomial parametrization (3.20) for u can be solved by primary decomposition. Every solution $(\xi_0, \ldots, \xi_h)$ yields a point of X^σ. Theorem 3.40 applies in this context.

Example 3.44. The above ideas can be applied to the independence model of three random variables of cardinalities 2, 3 and 3. As explained in the introduction, this is the smallest independence model to which the analytical solution of Ay and Knauf [7] does not apply.

The dimension of the model is $d = 5$ and the state space has cardinality 18, so the normal space $\mathcal{N}$ has dimension $18 - 5 - 1 = 12$. The symmetry group of the model is generated by the permutation of the two random variables of cardinality three and the permutations within the state spaces of each random variable.

The cocircuits can be computed by `TOPCOM`. Testing all 3^{18} possible sign vectors of length 18 shows that there are 365 592 nonzero sign vectors in $\mathcal{N}$ up to symmetry (this is the second algorithm proposed in A.2; the first algorithm takes too long in this case). Only 975 of them are quasi-critical. Excluding all sign vectors where the support of both the negative and the positive part exceeds 6 (cf. Lemma 3.5) reduces the problem to 240 sign vectors.

Again, the 72 sign vectors that do not have full support are easier to handle. They can be treated, for example, with the algorithm from the previous section. For the 168 sign vectors that have full support this is not possible. The corresponding systems of equations consist of $\dim \mathcal{N} - 1 = 11$ equations of $\dim \mathcal{N} = 12$ variables. The computer takes too long to solve these equations, but the sign vectors can be treated using the method proposed in this section, which only requires the primary decomposition of a

system of $d = 5$ polynomials in $d+1 = 6$ variables. The treatment of the sign vectors without full support is also faster with the algorithm from this section.

The analysis was carried out with the help of SINGULAR. It proved to be advantageous to use the algorithm of Shimoyama and Yokoyama (SY) implemented in the library `primdec.lib` [56]. The following result was obtained:

Theorem 3.45. *The maximal value of $D_{\mathcal{E}}$ for the independence model of cardinalities 2, 3 and 3 equals $\log(3+2\sqrt{2}) \approx 1.76$, and the maximal value of $\overline{D}_{\mathcal{E}}$ is $\log(2(1+\sqrt{2})) \approx 1.57$. Up to symmetry there is a unique global maximizing probability distribution*

$$(1 - \frac{\sqrt{2}}{2})(\delta_{122} + \delta_{111}) + (\sqrt{2} - 1)\delta_{000}.$$

In order to compare the two methods of finding the maximizers of $\overline{D}_{\mathcal{E}}$ and $D_{\mathcal{E}}$ presented in this section and in the last section let d be the dimension of the model and let $r = \dim \mathcal{N}$. Choose a sufficient statistics A such that $h = d + 1$. For any sign vector σ with full support, the algorithm from Section 3.6 starts with $r - 1$ equations (corresponding to a basis of $K_{\mathbb{C}}^{\sigma}$) in r variables $\lambda_1, \ldots, \lambda_r$. On the other hand, the algorithm in this section starts with the $d+1$ equations $Au = 0$ in the $d+2$ variables $\xi_0, \ldots, \xi_{d+1}$. Thus, heuristically, the first method should perform better when the codimension of the model is small, and the second method should perform better when the dimension of the model is small.

Remark 3.46. The discussion in this section also shows, that the projection property with respect to an algebraic exponential family is an "algebraic property," in the sense, that the projection points can be computed by solving polynomial equations. This fact was already noticed by Bernd Sturmfels, who proposed the following algorithm to compute the projection points (unpublished):

1. Let I be the ideal in the polynomial ring $\mathbb{C}[P(x), Q(x) : x \in \mathcal{X}]$ with $2N$ variables generated by the following polynomials:

$$\begin{aligned} Q^{v^+}\nu^{v^-} - Q^{v^-}\nu^{v^+}, &\qquad \text{for all } v \in \mathcal{N}, \\ AQ - AP, & \\ P(x)P(y)(P(x)Q(y) - P(y)Q(x)), &\qquad \text{for all } x, y \in \mathcal{X}. \end{aligned}$$

   The first equations are equivalent to $Q \in \overline{\mathcal{E}}$ by Theorem 2.32. The equation $AP = AQ$ implies that $Q = P_{\mathcal{E}}$. The last set of equations expresses the fact, that $\frac{P(x)}{P(y)} = \frac{Q(x)}{Q(y)}$, unless $P(x)P(y) = 0$.

2. Eliminate the variables $Q(x)$, i.e. compute $I \cap \mathbb{C}[P(x) : x \in \mathcal{X}]$.

3. Saturate with respect to the equations $P^{v^+}\nu^{v^-} - P^{v^-}\nu^{v^+}$ for all $v \in \mathcal{N}$, in order to remove the trivial solutions with $P \in \overline{\mathcal{E}}$.

Then the probability measures that satisfy the resulting ideal are the proper projection points.

One drawback of this algorithm is that the starting ideal I is an ideal in $2N$ variables. Even if the support of P is known (or if each possible support of P is considered separately), the starting ideal will be an ideal in more than N variables (unless $\text{supp}(P_{\mathcal{E}}) \neq \mathcal{X}$). In contrast to this, the algorithms proposed in Section 3.6 and in this section deal with ideals in polynomial rings with less than N variables. In fact, depending on the codimension of the model, one of the two algorithms will always work with a polynomial ring in at most $\frac{N+1}{2}$ variables.

4. Examples

The purpose of this chapter is to discuss some examples and to compare the two optimization problems of $D_{\mathcal{E}}$ and $\overline{D}_{\mathcal{E}}$. The theoretical results from the previous chapter are valuable tools in the study of the maximizers of $D_{\mathcal{E}}$, mainly because of two reasons:

1. The dimension of the problem is reduced.

Instead of maximizing over the whole probability simplex the maximization takes place over the set of kernel distributions $\mathrm{K}_{\mathcal{E}}$ or the bounded subset $\partial\mathbf{U}_{\mathcal{N}}$ of the normal space $\mathcal{N}$, respectively. Therefore, the dimension of the problem is reduced by the dimension of the exponential family. One has to admit that in view of Lemma 3.5 the gain is not so large for low-dimensional exponential families, see Section 4.1.

2. It is not necessary to compute an rI-projection onto $\mathcal{E}$ in order to evaluate $\overline{D}_{\mathcal{E}}$.

This is particularly important for the numerical search for maximizers using gradient search algorithms, which is now feasible for larger models. There may be many local maximizers and saddle points, however, so it is still a difficult problem to find the global maximizers of $D_{\mathcal{E}}$, see Example 4.29.

Theorem 3.28 and Lemmas 3.17 and 3.29 allow to translate results on the maximizers of one problem to the other problem. Therefore, it might seem irrelevant to discuss differences between the two optimization problems. There are, however, some differences between the two formulations. First, different points of view lead to different intuitions, and some results are easier to see in one formulation than in the other. Second, there is a more technical difference: Any probability measure $P \in \mathbf{P}(\mathcal{X})$ gives a lower bound of $\max D_{\mathcal{E}}$, if its rI-projection can be computed. In this case, $u = \Psi_{\mathcal{E}}(P)$ can also be computed and gives a lower bound of $\max \overline{D}_{\mathcal{E}}$. In the other direction, if a basis of $\mathcal{N}$ is known, then one can construct elements $u \in \partial\mathbf{U}_{\mathcal{N}}$, but since it is difficult to compute the rI-projection of u^+, it may not be possible to compute $D_{\mathcal{E}}(u^+)$ exactly.

Sections 3.6 and 3.7 demonstrated that the results of the previous chapter yield new algorithms for exact computations. The case where the exponential family $\mathcal{E}$ has codimension one, discussed in Examples 3.13 and 3.20, is another clear example of the elegance of the theory of maximizing $\overline{D}_{\mathcal{E}}$. Section 4.1 discusses the opposite case, when the exponential families has a low dimension. In this case the maximization of $D_{\mathcal{E}}$ and the maximization of $\overline{D}_{\mathcal{E}}$ are very similar. Section 4.2 introduces partition models, a class of convex exponential families with peculiar properties. For example, the smallest exponential families $\mathcal{E}$ with $\max D_{\mathcal{E}} = \log(2)$ are partition models, as Section 4.3 shows. Partition models also appear in the study of symmetric exponential families. Section 4.4 discusses binomial models and binary i.i.d. models, two classes of one-dimensional exponential families.

4.1. Low-dimensional exponential families

This section is devoted to low-dimensional exponential families. The trivial zero-dimensional case, in which both maximization problems are easy to understand, is summarized in 4.1.1. The main emphasis of this section is on one-dimensional exponential families, which are studied in 4.1.2. In 4.1.3 the results are specialized to the case where the cardinality of $\mathcal{X}$ equals four. In this case the number of local maximizers and their support sets can be computed as a function of the tangent space and the reference measure. The methods used in this section generalize and are also useful to handle other low-dimensional exponential families. This is sketched in 4.1.4. Two other classes of (closures of) one-dimensional exponential families, the binary i.i.d. models and the binomial models, will be treated later in Section 4.4.

4.1.1. Zero-dimensional exponential families

A zero-dimensional exponential family is just a single point $\mathcal{E} = \{\nu\}$ in the interior of $\mathbf{P}(\mathcal{X})$, which can be taken as the reference measure. Both the optimization problem of $D_{\mathcal{E}}$ and the maximization of $\overline{D}_{\mathcal{E}}$ are easy to solve. The normal space $\mathcal{N}$ is the orthogonal complement of $\mathbf{1}$ in $\mathbb{R}^{\mathcal{X}}$, i.e. $\mathcal{N}$ consists of all $u \in \mathbb{R}^{\mathcal{X}}$ such that $\sum_x u^+(x) = \sum_x u^-(x)$. Note that, if $\mathcal{E}'$ is an arbitrary exponential family with normal space $\mathcal{N}'$ and reference measure $\nu \in \mathbf{P}(\mathcal{X})$, then the maximization of $\overline{D}_{\mathcal{E}'}$ over $\partial \mathbf{U}_{\mathcal{N}'}$ can be formulated as the maximization of $\overline{D}_{\mathcal{E}}$ over $\partial \mathbf{U}_{\mathcal{N}}$ subject to linear constraints.

Any probability measure P satisfies $P_{\mathcal{E}} = \nu$. Therefore, P is a projection point if and only if P is a truncation of ν. By Proposition 2.14 (*iii*) the function $D_{\mathcal{E}} = D(\cdot\|\nu)$ is strictly convex, so every local maximizer is a point measure. Conversely, by Lemma 3.6 any point measure is a local maximizer with $D(\delta_i\|\mathcal{E}) = -\log(\nu_i)$. For any subset $\mathcal{Z} \subseteq \mathcal{X}$ the convex function $D_{\mathcal{E}}$ has a unique minimum in the interior of $\mathbf{P}(\mathcal{Z})$, and this minimum agrees with the unique projection point $P = \nu^{\mathcal{Z}}$ in $\mathbf{P}(\mathcal{Z})$.

The critical points of $\overline{D}_{\mathcal{E}}$ are of the form $\nu^{\mathcal{Z}} - \nu^{\mathcal{X}\setminus\mathcal{Z}}$ for nonempty subsets $\mathcal{Z} \subset \mathcal{X}$ such that $\mathcal{Z} \neq \mathcal{X}$. The value of $\overline{D}_{\mathcal{E}}$ at such a critical point is $\overline{D}_{\mathcal{E}}(\nu^{\mathcal{Z}} - \nu^{\mathcal{X}\setminus\mathcal{Z}}) = \log \frac{\nu(\mathcal{X}\setminus\mathcal{Z})}{\nu(\mathcal{Z})}$.

4.1.2. One-dimensional exponential families

Let $\mathcal{E}$ be a one-dimensional exponential family on $\mathcal{X} = \{1, \ldots, N\} = [N]$, where $N \geq 3$. Then $\mathcal{E}$ has a sufficient statistics matrix of the form

$$A = (a_1, a_2, \ldots, a_N)\,,$$

where $\{a_1, \ldots, a_N\}$ has cardinality at least two. Reordering $\mathcal{X}$ if necessary one may assume that $a_i \leq a_{i+1}$ for all i. Then $a_1 < a_N$. If $a_1 \neq 0$, then replace A by $A - a_1\mathbf{1}$. If $a_N \neq 1$, then replace A by $\frac{1}{a_N}A$. After these transformations the sufficient statistics is of the form

$$A = (0, a_2, \ldots, a_{N-1}, 1) \tag{4.1}$$

with $0 = a_1 \leq a_2 \leq \cdots \leq a_{N-1} \leq a_N = 1$.

Remark 4.1. There is one remaining symmetry: replacing v by $\mathbf{1} - v$ and reordering $\mathcal{X}$ replaces a_i and $(1 - a_{N+1-i})$. This symmetry can be used to reduce the number of cases in certain case distinctions.

For any $1 \leq i < j \leq N$ let $\Delta_{i,j}$ be the line segment between δ_i and δ_j. The relative interior of $\Delta_{i,j}$ is denoted by $\Delta^\circ_{i,j}$. The following theorem sums up the main results of the following calculations. These calculations contain much more information than stated in the theorem. For example, they contain a characterization of the maximizers in the convex case, and they contain information on other critical points. Furthermore, the calculations are important in their own right, since they demonstrate a general method for computing the local maximizers of low-dimension exponential families, as explained in Section 4.1.4.

Theorem 4.2. *Let $\mathcal{E}$ be a one-dimensional exponential family on $[N]$ with sufficient statistics given by (4.1). If the set $\{a_k : k = 1, \dots, N\}$ contains only two elements, then $\mathcal{E}$ is a convex set. Otherwise, the following holds:*

- *For any $i \in [N]$, if $a_1 < a_i < a_N$, then δ_i is a local maximizer.*
- *Let $i, j \in [N]$. If there is no a_k such that $a_i < a_k < a_j$, then there is no local maximizer in $\Delta^\circ_{i,j}$. Otherwise, there is at most one local maximizer in $\Delta^\circ_{i,j}$.*
- *If $a_i = 0$ and $a_j = 1$, then there is one local maximizer in $\Delta^\circ_{i,j}$.*

There are no further local maximizers. Hence, the total number of local maximizers is bounded from above by

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

If $0 < a_2 \leq \cdots \leq a_{N-1} < 1$, then the number of local maximizers is bounded from below by

$$(N-2) + 1 = N - 1.$$

Remark 4.3. The two bounds on the number of local maximizers are strict for $N = 4$, see Subsection 4.1.3, i.e. there are exponential families on $\mathcal{X} = \{1, 2, 3, 4\}$ with three and with five maximizers. Interestingly, there are no exponential families on $\mathcal{X}$ with four local maximizers.

It is more difficult to make general statements on the global maximizers. In many cases, the global maximizers are point measures, but not always. For example, the global maximizers for the binary i.i.d. models are point measures, and the global maximizers of the binomial models are supported on two states, as discussed in Section 4.4. For $N = 4$, any of the local maximizers that Theorem 4.2 allows may be the global maximizer, see Remark 4.7. This is not true if $\mathcal{E}$ is assumed to include the uniform distribution:

Proposition 4.4. *Let $\mathcal{E}$ be a one-dimensional exponential family on $[N]$ with uniform reference measure and sufficient statistics given by (4.1). Let $a_{\xi=1} = \frac{1}{N}\sum_i a_i$.*

(*i*) *If $a_k = a_{\xi=1}$ for some k, then the set of global maximizers of $D_{\mathcal{E}}$ consists of all point measures δ_i with $a_i = a_k$.*

(*ii*) *Otherwise, there exists $k < N$ such that $a_k < a_{\xi=1} < a_{k+1}$. If a global maximizer P of $D_{\mathcal{E}}$ has support $\{i, j\}$ for some $1 \leq i < j \leq N$, then either $a_i = 0$ and $a_j = a_{k+1}$, or $a_i = a_k$ and $a_j = 1$. In this case, $a_k < \sum_l P(l)a_l < a_{k+1}$. If δ_i is a global maximum, then $a_i \in \{a_k, a_{k+1}\}$.*

The proof of the proposition will be given at the end of this section. The statement of this proposition may not be optimal: In all known cases the global maximizer of the information divergence from a one-dimensional exponential family $\mathcal{E}$ containing the uniform distribution is a point measure, unless $\mathcal{E}$ is convex.

The calculations below involve functions of the form

$$f : \vartheta > 0 \mapsto \sum_{k=0}^{n} \nu_k \vartheta^{b_k},$$

where $\nu_k, b_k \in \mathbb{R}$ for all k. One may assume that $\nu_k \neq 0$ for all k and that $b_0 < b_1 < \cdots < b_n$. If all b_k are integral, then f is a Laurent polynomial, or even a polynomial if all b_k are non-negative. If all b_k are rational, then f is a Puiseux polynomial. In the general case f can be identified with a finite Hahn series over $\mathbb{R}$ with value group $\mathbb{R}$ (one might be tempted to call such a function a *Hahn polynomial*, but this name is already reserved for a class of orthogonal polynomials).

In particular it will be important to count sign changes of f in an interval $0 \leq \vartheta_{\min} < \vartheta < \vartheta_{\max} \leq \infty$. Since f is real analytic for $\vartheta > 0$, the zeros of f are isolated, unless $f = 0$. Furthermore, for large ϑ the highest order term $\nu_l \vartheta^{b_l}$ with $l = \operatorname{argmax}\{b_k : \nu_k \neq 0\}$ dominates, therefore the number of sign changes is finite. Only "true" sign changes from + to − or from − to + are counted. For example, the function $f(\vartheta) = (\vartheta - 1)^2$ has no sign changes, and the finite sequence $1, 0, 1, 0, -1$ has only one sign change.

The following lemma is well-known for polynomials under the name of *Descartes' rule of signs*. It formalizes the following intuition: For small ϑ the lowest order term determines the behaviour of f, and the highest order term determines the behaviour for large ϑ. For intermediate ranges of ϑ, the intermediate terms play a role, one after the other. The polynomial version of the lemma is usually proved using induction and polynomial division. Unfortunately, the division algorithm does not work here.

Lemma 4.5. *Let $f(\vartheta) = \sum_{k=0}^{n} \nu_k \vartheta^{b_k}$, where $\nu_k, b_k \in \mathbb{R}$ and $b_0 < b_1 < \cdots < b_n$. Let c_f be the number of sign changes of the function f on $\mathbb{R}_{\geq}$, and let c_ν be the number of sign changes in the sequence $\nu_0, \nu_1, \ldots, \nu_n$. Then $c_f \leq c_\nu$, and $c_\nu - c_f$ is even.*

Proof. Suppose that none of the coefficients ν_k, $k = 0, \ldots, n$ vanishes. Then one may divide by ν_n and assume without loss of generality that $\nu_n = 1$. With this normalization $\lim_{\vartheta \to +\infty} f(\vartheta) = +\infty$. If $\nu_0 < 0$, then $f(\vartheta) < 0$ for sufficiently small $\vartheta > 0$, so c_f is odd, otherwise, if $\nu_0 \geq 0$, then c_f is even. This proves that $c_\nu - c_f$ is even.

The inequality $c_f \leq c_\nu$ follows from induction on n: If $n = 1$, then f is monotone, and f has a root if and only if $a_0 < 0$, so the statement holds. Suppose that the statement holds for all such functions with n or less terms.

Assume that $b_0 = 0$ and $\nu_0 \neq 0$. This is possible without loss of generality, since the numbers of sign changes of f and of $\vartheta^{-\nu_0} f$ are the same. Let

$$g = \vartheta \frac{\partial}{\partial \vartheta} f = \sum_{k=1}^{n} b_k \nu_k \vartheta^{b_k}.$$

Then g has less terms than f. Let c'_ν be the number of sign changes in the sequence $b_1\nu_1, \ldots, b_n\nu_n$, and let c_g be the number of sign changes of the function g on $\mathbb{R}_{\geq}$. By the induction assumption, $c_g \leq c'_\nu$. Clearly $c'_\nu \leq c_\nu$.

Assume that the function f changes its sign at ϑ_0. Then g cannot change its sign at ϑ_0, for otherwise f would have a local extremum at ϑ_0. Let $\vartheta_1, \ldots, \vartheta_{c_f}$ be the places where f changes its sign. Then $\frac{\partial}{\partial \vartheta} f$ must change its sign in each open interval $(\vartheta_i, \vartheta_{i+1})$. Hence the same is true for g, so $c_g \geq c_f - 1$. Therefore,

$$c_f - 1 \leq c_g \leq c'_\nu \leq c_\nu.$$

Since c_f and c_ν are either both even or both odd, it follows that $c_f \leq c_\nu$ □

The calculations proving Theorem 4.2 is spread across four paragraphs: First, the convex case is discussed. Second, assuming that $\mathcal{E}$ is not convex, certain right inverses of the rI-projection map are computed. This is useful, since there is no analytic formula for the rI-projection itself. In a third step, these inverses can be used to compute the derivatives of the information divergence. Finally, an analysis of the derivatives yields the statements of Theorem 4.2. It is also possible to do the same calculations for the maximization of $\overline{D}_\mathcal{E}$; this will be sketched at the end of the calculations.

The one-dimensional exponential family with sufficient statistics (4.1) and reference measure ν can be parametrized by

$$P_\xi(i) = \frac{\nu_i}{Z_\xi} \xi^{a_i}, \qquad \text{for all } i \in \mathcal{X},$$

where $0 < \xi < \infty$. The convex support is the closed one-dimensional interval $[0, 1]$. By Lemma 3.5 the support of any local maximizer has cardinality one or two. Let $i_0 < j_0$ be such that $0 = a_1 = \cdots = a_{i_0} < a_{i_0+1}$ and $a_{j_0-1} < a_{j_0} = \cdots = a_N = 1$. Let $\mathcal{Z} = \{1, \ldots, i_0\}$ and $\mathcal{Z}' = \{j_0, \ldots, N\}$.

The convex case. Assume that $j_0 = i_0 + 1$. Then $\overline{\mathcal{E}}$ is actually a linear family: $\overline{\mathcal{E}}$ consists of all probability measures of the form $P_\lambda = (1 - \lambda)\nu^\mathcal{Z} + \lambda\nu^{\mathcal{Z}'}$. In particular, $\mathcal{E}$ is a convex exponential family. Such exponential families and their maximizers have been studied by Ay and Matúš in [52]. If $\mathbf{1}$ is a reference measure, then $\overline{\mathcal{E}}$ also belongs to the class of partition models, see Section 4.2.

For any probability measure P the moment map computes $\pi_A(P) = P(\mathcal{Z})$. Therefore, $P_\mathcal{E} = P_{\lambda(P)}$, where $\lambda(P) = P(\mathcal{Z}')$. Write $P = (1 - \lambda(P))P^\mathcal{Z} + \lambda(P)P^{\mathcal{Z}'}$. Then

$D(P\|\mathcal{E}) = P(\mathcal{Z})D(P^{\mathcal{Z}}\|\nu^{\mathcal{Z}}) + P(\mathcal{Z}')D(P^{\mathcal{Z}'}\|\nu^{\mathcal{Z}'})$. Without loss of generality assume that $\max\{D(P\|\mathcal{E}) : P \in \mathbf{P}(\mathcal{Z})\} \geq \max\{D(P\|\mathcal{E}) : P \in \mathbf{P}(\mathcal{Z}')\} =: C$. Then the set of all local maximizers equals the set

$$\begin{aligned}&\left\{\delta_i : i \in \mathcal{Z}, D(\delta_i\|\nu^{\mathcal{Z}}) > C\right\}\\ &\qquad \cup \left\{(1-\lambda)\delta_i + \lambda\delta_j : i \in \mathcal{Z}, j \in \mathcal{Z}, D(\delta_i\|\nu^{\mathcal{Z}}) = D(\delta_j\|\nu^{\mathcal{Z}'}) = C, 0 \leq \lambda < 1\right\}.\end{aligned}$$

This set is connected if and only if either $\max\{D(P\|\mathcal{E}) : P \in \mathbf{P}(\mathcal{Z})\} = \max\{D(P\|\mathcal{E}) : P \in \mathbf{P}(\mathcal{Z}')\}$ or if the set $\left\{\delta_i : i \in \mathcal{Z}, D(\delta_i\|\nu^{\mathcal{Z}}) \geq C\right\}$ consists of a single element.

If P is a probability measure with support contained in $\mathcal{Z}$, then the support of its rI-projection $P_{\mathcal{E}}$ equals $\mathcal{Z}$ by Lemma 2.28 (in fact, $P_{\mathcal{E}} = \nu^{\mathcal{Z}}$). If P is a local maximizer of $D_{\mathcal{E}}$, then $\{A_x : x \in \mathcal{Z}\}$ and $\{A_x : x \in \mathcal{X}\backslash\mathcal{Z}\}$ must lie in distinct parallel hyperplanes by Theorem 3.2 (i). Hence $j_0 = i_0 + 1$, and P is a point measure by Lemma 3.5. The same statements hold true on $\mathcal{Z}'$. Therefore, if $\overline{\mathcal{E}}$ is not convex, then there are no local maximizers with support contained in $\mathcal{Z}$ or $\mathcal{Z}'$.

Right inverses of the rI-projection. In the following assume that $j_0 > i_0 + 1$ and let $\mathcal{Y} = \{i_0 + 1, \ldots, j_0 - 1\}$. All point measures δ_i for $i \in \mathcal{Y}$ are local maximizers by Lemma 3.6. All other possible local maximizers have a support of cardinality 2. Each such maximizer lies in the convex hull $\Delta_{i,j}$ of two point measures δ_i, δ_j, and $a_i \neq a_j$ by Lemma 3.5.

Fix $1 \leq i < j \leq N$ such that $a_i < a_j$. Any probability measure on $\Delta_{i,j}$ is of the form

$$P_\lambda = (1-\lambda)\delta_i + \lambda\delta_j.$$

While it is difficult to compute the mapping $\lambda \mapsto \xi(\lambda)$ such that $(P_\lambda)_{\mathcal{E}} = P_{\xi(\lambda)}$, the inverse mapping $\xi \mapsto \lambda(\xi)$ is easy to obtain: Observe that

$$\pi_A(P_\lambda) = (1-\lambda)a_i + \lambda a_j \qquad \text{and} \qquad \pi_A(P_\xi) = \frac{1}{Z_\xi}\sum_{k\in\mathcal{X}} a_k\nu_k\xi^{a_k}.$$

From $a_\xi := \pi_A(P_\xi) = \pi_A(P_{\lambda(\xi)})$ it follows that

$$\lambda(\xi) = \frac{a_\xi - a_i}{a_j - a_i} = \frac{\sum_{k\in\mathcal{X}}(a_k - a_i)\nu_k\xi^{a_k}}{(a_j - a_i)Z_\xi}. \tag{4.2}$$

There exist $\xi_i \geq 0$ and $\xi_j \leq \infty$ such that $\lambda \to 0$ for $\xi \to \xi_i$ and $\lambda \to 1$ for $\xi \to \xi_j$. Then $\xi_i = 0$ if and only if $i \in \mathcal{Z}$, and $\xi_j = \infty$ if and only if $j \in \mathcal{Z}'$.

The information divergence and its derivatives. $D(P_\lambda\|\mathcal{E})$ equals

$$D(P_\lambda\|P_{\xi(\lambda)}) = (1-\lambda)\log\frac{1-\lambda}{\nu_i\xi^{a_i}} + \lambda\log\frac{\lambda}{\nu_j\xi^{a_j}} + \log Z_\xi. \tag{4.3}$$

By Theorem 3.1 the directional derivative with respect to λ equals

$$\frac{\partial}{\partial\lambda}D(P_\lambda\|P_{\xi(\lambda)}) = -\log\frac{P_\lambda(i)}{P_\xi(i)} + \log\frac{P_\lambda(j)}{P_\xi(j)} = \log\frac{\nu_i\lambda}{\nu_j(1-\lambda)\xi^{a_j-a_i}}. \tag{4.4}$$

If $i \notin \mathcal{Z}$, then $\frac{\partial}{\partial\lambda}D(P_\lambda\|P_{\xi(\lambda)}) \to -\infty$ for $\xi \to \xi_i$. Similarly, if $j \notin \mathcal{Z}'$, then $\frac{\partial}{\partial\lambda}D(P_\lambda\|P_{\xi(\lambda)}) \to +\infty$ for $\xi \to \xi_j$. This expresses the fact, that all point measures δ_k for $k \in \mathcal{Y}$ are local maximizers. Otherwise, assume that $i \in \mathcal{Z}$. Then $a_i = 0$, so

$$\begin{aligned}\lim_{\lambda\to 0}\frac{\partial}{\partial\lambda}D(P_\lambda\|P_{\xi(\lambda)}) &= \lim_{\xi\to 0}\log\frac{\nu_i\lambda(\xi)}{\nu_j\xi^{a_j}} = \log\frac{\nu_i}{\nu_j a_j Z_0} + \lim_{\xi\to 0}\log\sum_{k>i_0} a_k\nu_k\xi^{a_k-a_j},\\ &= \begin{cases}\log\frac{\nu_i\sum_{k:a_k=a_{i_0+1}}\nu_k a_k}{\nu_j a_j Z_0}, & \text{if } a_j = a_{i_0+1},\\ +\infty, & \text{else.}\end{cases}\end{aligned} \tag{4.5}$$

where $Z_0 = \sum_{k\in\mathcal{Z}}\nu_k$.

The second derivative of $D(P_\lambda\|P_{\xi(\lambda)})$ with respect to λ equals

$$\frac{\partial^2}{\partial\lambda^2}D(P_\lambda\|P_{\xi(\lambda)}) = \frac{1}{\lambda} + \frac{1}{1-\lambda} - \frac{a_j-a_i}{\xi}\frac{\partial\xi}{\partial\lambda} = \frac{1}{\lambda(1-\lambda)} - \frac{a_j-a_i}{\xi}\frac{\partial\xi}{\partial\lambda}.$$

Using

$$1-\lambda = \frac{a_j-a_\xi}{a_j-a_i} = \frac{\sum_{k\in\mathcal{X}}(a_j-a_k)\nu_k\xi^{a_k}}{(a_j-a_i)Z_\xi}$$

and

$$\begin{aligned}\frac{\partial\lambda}{\partial\xi} &= \frac{\sum_{k,l}(a_k-a_i)a_k\nu_k\nu_l\xi^{a_k+a_l} - \sum_{k,l}(a_k-a_i)a_l\nu_k\nu_l\xi^{a_k+a_l}}{(a_j-a_i)\xi Z_\xi^2}\\ &= \frac{\sum_{k,l}(a_k-a_l)a_k\nu_k\nu_l\xi^{a_k+a_l}}{(a_j-a_i)\xi Z_\xi^2}\end{aligned}$$

this rewrites to

$$\begin{aligned}\frac{\partial^2}{\partial\lambda^2}D(P_\lambda\|P_\xi) &= Z_\xi^2(a_j-a_i)^2\left(\frac{1}{\sum_{k,l}(a_k-a_i)(a_j-a_l)\nu_k\nu_l\xi^{a_k+a_l}}\right.\\ &\qquad\left. - \frac{1}{\sum_{k,l}(a_k-a_l)a_k\nu_k\nu_l\xi^{a_k+a_l}}\right)\\ &= \frac{C(\xi)}{Z(\xi)^2}\left(\sum_{k,l}(a_k-a_l)a_k\nu_k\nu_l\xi^{a_k+a_l} + \sum_{k,l}(a_k-a_i)(a_l-a_j)\nu_k\nu_l\xi^{a_k+a_l}\right)\\ &= \frac{C(\xi)}{Z(\xi)}\sum_k(a_k-a_i)(a_k-a_j)\nu_k\xi^{a_k},\end{aligned} \tag{4.6}$$

where

$$C(\xi) = \frac{(a_j - a_i)^2 Z_\xi^2}{(a_\xi - a_i)(a_j - a_\xi)(\sum_{k,l}(a_k - a_l)a_k\nu_k\nu_l\xi^{a_k+a_l})}.$$

The variance of the function $k \mapsto a_k$ under the probability distribution P_ξ equals

$$\sigma^2(\xi) = \frac{1}{Z_\xi}\sum_k a_k^2\nu_k\xi^{a_k} - a_\xi^2 \geq 0;$$

it vanishes only in the limits $\xi \to 0$ and $\xi \to +\infty$. So if $\xi_i < \xi < \xi_j$, then $C(\xi) = \frac{(a_j-a_i)^2}{(a_\xi-a_i)(a_j-a_\xi)\sigma^2(\xi)} > 0$.

Assume that $a_i > 0$. Then $\xi_i > 0$, so $\frac{1}{Z_{\xi_i}}\sum_k(a_k - a_i)(a_k - a_j)\nu_k\xi_i^{a_k} = \sigma^2(\xi_i)$ is strictly positive. Therefore, $\frac{\partial^2}{\partial\lambda^2}D(P_\lambda\|P_{\xi(\lambda)}) > 0$ if ξ is close to ξ_i. Similarly, if $a_j < 1$, then $\frac{\partial^2}{\partial\lambda^2}D(P_\lambda\|P_{\xi(\lambda)}) > 0$ if ξ is close to ξ_j.

Analysis of the derivatives. If there is no a_k such that $a_i < a_k < a_j$, then the second derivative $\frac{\partial^2}{\partial\lambda^2}D(P_\lambda\|P_{\xi(\lambda)})$ is strictly positive, so $D_\mathcal{E}$ is strictly convex on $\Delta_{i,j}$. If $a_i = 0$ and $a_j = 1$, then $\frac{\partial^2}{\partial\lambda^2}D(P_\lambda\|P_{\xi(\lambda)})$ is strictly negative, so $D_\mathcal{E}$ is strictly concave on $\Delta_{i,j}$. If either $a_i = 0$ or $a_j = 1$, but not both, then Lemma 4.5 shows that $\frac{\partial^2}{\partial\lambda^2}D(P_\lambda\|P_{\xi(\lambda)})$ has up to one sign change. Hence $\frac{\partial}{\partial\lambda}D(P_\lambda\|P_{\xi(\lambda)})$ has up to two zeros, corresponding to up to two critical points of $D_\mathcal{E}$ on the relative interior $\Delta_{i,j}^\circ$ of $\Delta_{i,j}$, and at most one of them may be a local maximizer. Assume that $0 < a_i < a_j < 1$. Lemma 4.5 shows that $\frac{\partial^2}{\partial\lambda^2}D(P_\lambda\|P_{\xi(\lambda)})$ has up to two sign changes for $\xi_i < \xi < \xi_j$ and may be negative in a subinterval of $[\xi_i, \xi_j]$. Therefore, $\frac{\partial}{\partial\lambda}D(P_\lambda\|P_{\xi(\lambda)})$ has up to three zeros, corresponding to up to three critical points of $D_\mathcal{E}$ in $\Delta_{i,j}^\circ$, and only one of them may be a local maximizer, since the boundary points δ_i and δ_j are also local maximizers of $D_\mathcal{E}$.

From this differential analysis, the following general statements can be deduced:

- If $i \in \mathcal{Z}$ and $j \in \mathcal{Z}'$, then $D_\mathcal{E}$ is concave on $\Delta_{i,j}$. Since the directional derivative of $D_\mathcal{E}$ along $\Delta_{i,j}$ is $+\infty$ at δ_i and $-\infty$ at δ_j, there is exactly one local maximizer with support $\{i, j\}$.

- Assume $i \in \mathcal{Z}$ and $j \in \mathcal{Y}$. Let $\delta = \lim_{\xi\to 0}\frac{\partial}{\partial\lambda}D(P_\lambda\|P_{\xi(\lambda)})$. If $a_j = a_{i_0+1}$, then $D_\mathcal{E}$ is convex on $\Delta_{i,j}$. In this case, if $\delta < 0$, then the restriction $D_\mathcal{E}|_{\Delta_{i,j}}$ of $D_\mathcal{E}$ to $\Delta_{i,j}$ has a local minimum in $\Delta_{i,j}^\circ$, and if $\delta \geq 0$, then there is no critical point in $\Delta_{i,j}^\circ$.

  Otherwise, if $a_j > a_{i_0+1}$, then $\delta = +\infty$ by (4.5). There may be up to two critical points in $\Delta_{i,j}^\circ$, and one of them may be a local maximizer.

- Similar remarks apply to the case $i \in \mathcal{Y}$, $j \in \mathcal{Z}'$.

- If $i, j \in \mathcal{Y}$, then the two endpoints δ_i and δ_j are local maximizers of $D_\mathcal{E}$. Hence there must be at least one local minimizer of $D_\mathcal{E}|_{\Delta_{i,j}}$ in $\Delta_{i,j}^\circ$, and the number of such minimizers is one more than the number of local maximizers in $\Delta_{i,j}^\circ$. If

there is no k such that $a_i < a_k < a_j$, then $D_{\mathcal{E}}$ is convex on $\Delta_{i,j}$, so there is no local maximizer in $\Delta_{i,j}^\circ$ in this case. Otherwise, Lemma 4.5 shows that the second derivative of $D_{\mathcal{E}}$ along $\Delta_{i,j}$ has at most two sign changes. Therefore, there are at most three critical points in $\Delta_{i,j}^\circ$, corresponding to at most one local maximizer.

For the proof of the theorem it remains to show the two bounds. The lower bound corresponds to the points measures δ_i for $i \in \mathcal{Y}$. Let $j_0' = N + 1 - j_0$. From the above statements it follows that the number of local maximizers is bounded from above by

$$\begin{aligned}(N - i_0 - j_0') + i_0 j_0' + i_0(N - i_0 - j_0' - 1) + j_0'(N - i_0 - j_0' - 1) \\ + \left(\binom{N - i_0 - j_0'}{2} - (N - i_0 - j_0' - 1)\right) \\ = \frac{1}{2}(N^2 - N - i_0^2 - j_0'^2 - i_0 - j_0'),\end{aligned}$$

and this value is maximal if $i_0 = 1$ and $j_0 = N$.

Maximizing $\overline{D}_{\mathcal{E}}$. Theorem 3.28 allows to translate the statements of Theorem 4.2 to statements about the local maximizers of $\overline{D}_{\mathcal{E}}$. It is instructive to derive the corresponding statements directly and to compare the necessary arguments. The calculations are similar; hence they will only be sketched.

It is easy to calculate the circuits of $\mathcal{N}$: If $a_i = a_j$ for some $i < j$, then (i, j) is a circuit, corresponding to the circuit vector $\delta_i - \delta_j$. If $a_i < a_j < a_k$ for some $i < j < k$, then (i, j, k) is a circuit, corresponding to the circuit vector $(a_k - a_j)\delta_i - (a_k - a_i)\delta_j + (a_j - a_i)\delta_k$. These are all the circuits: This expresses the fact that two columns i, j of A are affinely dependent if and only if $a_i = a_j$, and all triples of columns of A are affinely dependent.

From these considerations it is easy to determine all sign vectors of $\mathcal{N}$: A sign vector $\sigma \in \{0, \pm 1\}^{\mathcal{X}}$ occurs in $\mathcal{N}$ if and only if there are at least two true sign changes in the sequence $\sigma_1, \sigma_2, \ldots, \sigma_N$. To see this, note that in this case there exist $i < j < k$ such that $\sigma_i = \sigma_k = -\sigma_j \neq 0$, corresponding to an oriented circuit τ_0. Then there are signed circuits τ_l such that $\operatorname{supp}(\tau_l) \subseteq \{i, j, k, l\}$ and $(\tau_l)_l = \sigma_l$ for all $l = 1, \ldots, N$, whence $\sigma = \tau_0 \circ \tau_1 \circ \cdots \circ \tau_N$.

By Lemma 3.30 it suffices to consider the case that the support of σ^+ has cardinality one or two. If $\mathcal{Y} = \emptyset$, then all circuits have cardinality two and are contained either in $\mathcal{Z}$ or in $\mathcal{Z}'$. Therefore, the same statements as above can be concluded from Corollary 2.27.

Assume that $\mathcal{Y} \neq \emptyset$ in the following. For any local maximizer $u \in \partial \mathbf{U}_{\mathcal{N}}$ the support of u^+ cannot be contained in $\mathcal{Z}$, because of Proposition 3.21 (i) and Lemma 3.24. For any $i \in \mathcal{Y}$ there exists a local maximizer of $\overline{D}_{\mathcal{E}}$ such that $\operatorname{supp}(u^+) = \{i\}$: Since $\sigma = \delta_i - \sum_{j \neq i} \delta_j$ is a sign vector of $\mathcal{N}$ there exists $v \in \partial \mathbf{U}_{\mathcal{N}}$ such that $v_i = 1$ and $v_j < 0$ for all $j \neq i$. Let $u^- = \Phi(v^+)$, where Φ is defined as in Remark 3.27. In the notation of Lemma 3.26, u^- is the rI-projection of v^- to $\mathcal{E}^{\mathcal{X} \setminus \{i\}}$. By assumption

$\mathrm{supp}(v^-) = \mathcal{X} \setminus \{i\}$, and so u^- must also have support $\mathcal{X} \setminus \{i\}$. Then $u := \delta_i - u^-$ is a local maximizer of $\overline{D}_{\mathcal{E}}$ on $\partial \mathbf{U}_{\mathcal{N}}$, since u^- is a local minimizer of $D(\cdot \| \nu|_{\mathcal{X}\setminus\{i\}})$ and since there is a neighbourhood U of u in $\partial \mathbf{U}_{\mathcal{N}}$ such that any $v \in U$ has the same sign vector as u.

It remains to discuss the case that u^+ is contained in the relative interior $\Delta_{i,j}^\circ$ of $\Delta_{i,j}$ for some $i < j$ satisfying $a_i < a_j$. Let $u_\lambda^+ = (1-\lambda)\delta_i + \lambda\delta_j$. By Remark 3.27, if u_λ^+ is the positive part of some critical point, then the corresponding negative part is of exponential form $\Phi(u_\lambda^+)$. Let $\mathcal{X}' = \mathcal{X} \setminus \{i,j\}$, and let u_ζ^- be the probability measure with

$$u_\zeta^-(k) = \begin{cases} \frac{\nu_k}{Z'_\zeta}\zeta^{a_k}, & \text{if } k \in \mathcal{X}', \\ 0, & \text{else,} \end{cases}$$

where $Z'_\zeta = \sum_{k\in\mathcal{X}'} \nu_k \zeta^{a_k}$. Then $u_\lambda^+ - u_\zeta^- \in \mathcal{N}$ if and only if

$$\lambda = \lambda(\zeta) := \frac{\sum_{k\in\mathcal{X}'}(a_k - a_i)\nu_k\zeta^{a_k}}{(a_j - a_i)Z'_\zeta},$$

cf. (4.2). The function $\zeta \mapsto \lambda(\zeta)$ is injective, but in general it is difficult to compute the function $\lambda \mapsto \zeta(\lambda)$. It may happen that the image of the map $\zeta \mapsto \lambda(\zeta)$ is not the complete interval $(0,1)$. Namely, if for some $0 < \lambda < 1$ the probability measure u_λ^+ is not a kernel distribution, then there is no ζ such that $\lambda = \lambda(\zeta)$. This happens if $i = 1$ and $a_2 > 0$ or if $j = N$ and $a_{N-1} < 1$. Otherwise, there exist $\zeta_i \geq 0$ and $\zeta_j \leq +\infty$ such that $\zeta \mapsto \lambda(\zeta)$ maps the interval (ζ_i, ζ_j) onto the interval $(0,1)$.

Let $u_\lambda = u_\lambda^+ - u_{\zeta(\lambda)}^-$. Then

$$\overline{D}_{\mathcal{E}}(u_\lambda) = (1-\lambda)\log\frac{1-\lambda}{\nu_i} + \lambda\log\frac{\lambda}{\nu_j} - a_\zeta\log\zeta + \log Z_\zeta$$

has the same form as $D_{\mathcal{E}}(P_\lambda)$, cf. (4.3). The partial derivative of $\overline{D}_{\mathcal{E}}(u_\lambda)$ equals

$$\frac{\partial}{\partial\lambda}\overline{D}_{\mathcal{E}}(u_\lambda) = \log\frac{\nu_i\lambda}{\nu_j(1-\lambda)\zeta^{a_j-a_i}},$$

cf. (4.4). The asymptotic behaviour is similar to the asymptotic behaviour of (4.4), see (4.5). The second derivative of $\overline{D}_{\mathcal{E}}(u_\lambda)$ equals

$$\frac{\partial^2}{\partial\lambda^2}\overline{D}_{\mathcal{E}}(u_\lambda) = C'(\zeta)\sum_{k\in\mathcal{X}'}(a_k - a_i)(a_k - a_j)\nu_k\zeta^{a_k} = C'(\zeta)\sum_{k\in\mathcal{X}}(a_k - a_i)(a_k - a_j)\nu_k\zeta^{a_k},$$

where $C'(\zeta) > 0$, cf. (4.6).

This leads to the following peculiar situation: The two functions $D_{\mathcal{E}}(P_\lambda)$ and $\overline{D}_{\mathcal{E}}(u_\lambda)$ for $0 < \lambda < 1$ have the same critical points, and their second derivatives differ only by a positive factor $C'(\zeta)/C(\xi)$.

Proof of Proposition 4.4. The entropy of the uniform distribution is maximal. Therefore, the entropy at P_ξ

$$H(\xi) := -\sum_l P_\xi(l) \log P_\xi(l) = \log(Z_\xi) - a_\xi \log(\xi)$$

has a maximum at $\xi = 1$, where the notation from the calculations proving Theorem 4.2 was used. By (2.5), if $P \in \mathbf{P}(\mathcal{X})$ rI-projects to P_ξ, then $D_\mathcal{E}(P) \leq H(\xi) \leq H(1) = \log|\mathcal{X}|$, with equality if and only if P is a point measure projecting to the uniform distribution. This proves (i).

The function $\xi \mapsto a_\xi$ is monotone, and so the derivative $\frac{\partial a_\xi}{\partial \xi}$ is positive. From

$$\frac{\partial}{\partial \xi} H(\xi) = -\frac{\partial a_\xi}{\partial \xi} \log(\xi)$$

it follows that $H(\xi)$ is monotonically increasing for $\xi < 1$ and monotonically decreasing for $\xi > 1$. Let ξ_k, ξ_{k+1} such that P_{ξ_k} and $P_{\xi_{k+1}}$ equal the rI-projections of δ_k and δ_{k+1}, respectively (if $a_k = 0$, then let $\xi_k = 0$, and if $a_{k+1} = 1$, then let $\xi_{k+1} = \infty$).

Let $P \in \mathbf{P}(\mathcal{X})$ be a global maximizer of $D_\mathcal{E}$, and let P_ξ be its rI-projection. Then $H(\xi) \geq H(\xi) - H(P) = D_\mathcal{E}(P) \geq D_\mathcal{E}(\delta_k) = H(\xi_k)$. Similarly, $H(\xi) \geq H(\xi_{k+1})$, whence $\xi_k \leq \xi \leq \xi_{k+1}$. Therefore, if $P = \delta_i$ is a point measure, then $a_i \in \{a_k, a_{k+1}\}$. Otherwise, if P is not a point measure, then $\xi_k < \xi < \xi_{k+1}$, and so $a_k < a_\xi = \sum_l P(l) a_l < a_{k+1}$. Write $P = (1-\lambda)\delta_i + \lambda\delta_j$, where $\lambda = \frac{a_\xi - a_i}{a_j - a_i} \notin \{0, 1\}$. Applying the symmetry of Remark 4.1 one may assume that $\lambda \geq \frac{1}{2}$. Let $\lambda' = \frac{a_\xi}{a_l}$. Then $\lambda' \geq \lambda$, with equality if and only if $a_i = 0$ and $a_j = a_l$. The rI-projection of $P' := (1-\lambda')\delta_1 + \lambda'\delta_{k+1}$ agrees with P_ξ. Equation (2.5) specializes to

$$D_\mathcal{E}(P) = H(P_\xi) - H(P) = H(P_\xi) - h(\lambda, 1-\lambda),$$

where $h(\lambda, 1-\lambda)$ equals the entropy of a binary random variable. Similarly, $D_\mathcal{E}(P') = H(P_\xi) - h(\lambda', 1-\lambda')$. Hence $D_\mathcal{E}(P') \geq D_\mathcal{E}(P)$ follows from $\frac{1}{2} \leq \lambda \leq \lambda'$ and the fact that $h(\lambda, 1-\lambda)$ is concave in λ, with maximum at $\frac{1}{2}$. Therefore, $\lambda = \lambda'$, so $a_i = 0$ and $a_j = a_{k+1}$. □

4.1.3. One-dimensional exponential families on four states

In this section the results are specialized to the case $N = 4$. This gives an illustration, how the general statements of the previous section can look like in a concrete case. Let $\mathcal{E}$ be a one-dimensional exponential family on $\mathcal{X} = \{1, 2, 3, 4\}$. The discussion in the previous section shows that one may assume that

$$A = (0, s, t, 1) \tag{4.7}$$

is a sufficient statistics of $\mathcal{E}$, where $0 \leq s \leq t \leq 1$. In this case the symmetry of Remark 4.1 is

$$\begin{aligned} &1 \leftrightarrow 4, && 2 \leftrightarrow 3, \\ &s \leftrightarrow 1-t, && t \leftrightarrow 1-s. \end{aligned} \tag{4.8}$$

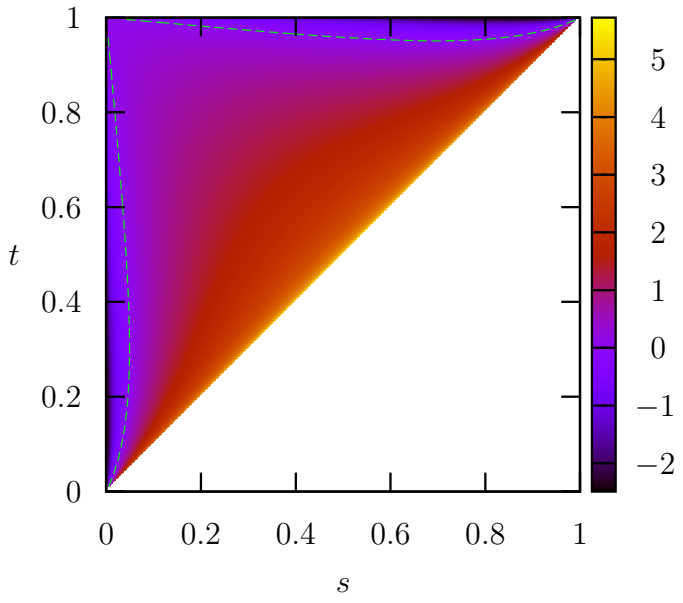

Figure 4.1.: A heat map of δ_ν for $\nu = \mathbf{1}$. For most values of s and t there are only six extremal points of $\overline{D}_{\mathcal{E}}$. If $\delta_\nu < 0$, then there are four more extremal points of $\overline{D}_{\mathcal{E}}$. This happens on the left and on the top, i.e., for small s or for large t. The cyan line marks the zero set of δ_ν.

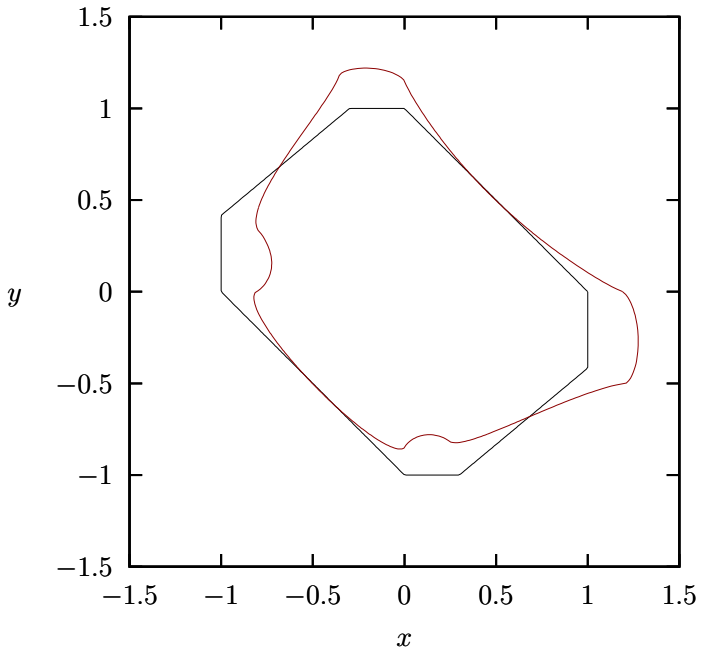

Figure 4.2.: The function $\overline{D}_{\mathcal{E}}$ for $s = \frac{1}{3}$ and $t = \frac{4}{5}$ as a polar plot. The six extreme points are clearly visible.

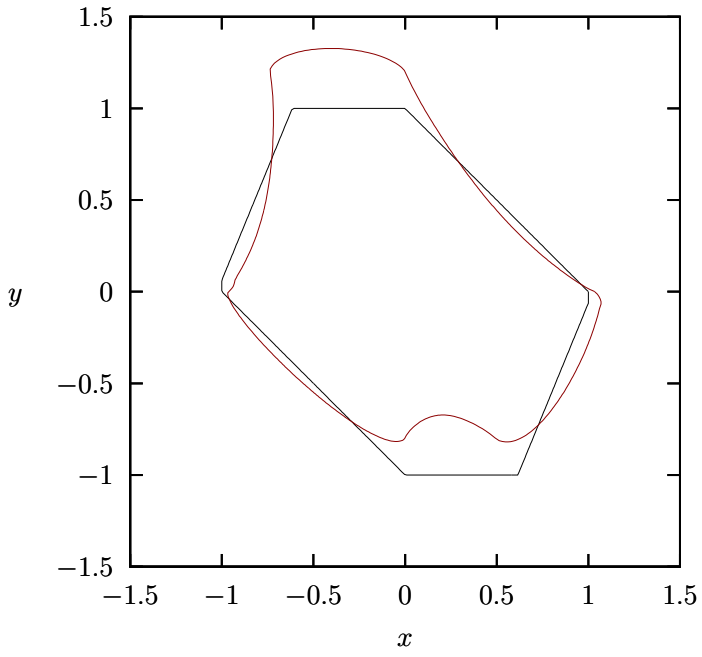

Figure 4.3.: The function $\overline{D}_{\mathcal{E}}$ for $s = \frac{1}{40}$ and $t = \frac{2}{5}$ as a polar plot. In this case the function $\overline{D}_{\mathcal{E}}$ has ten extreme points, but four of them are hardly visible.

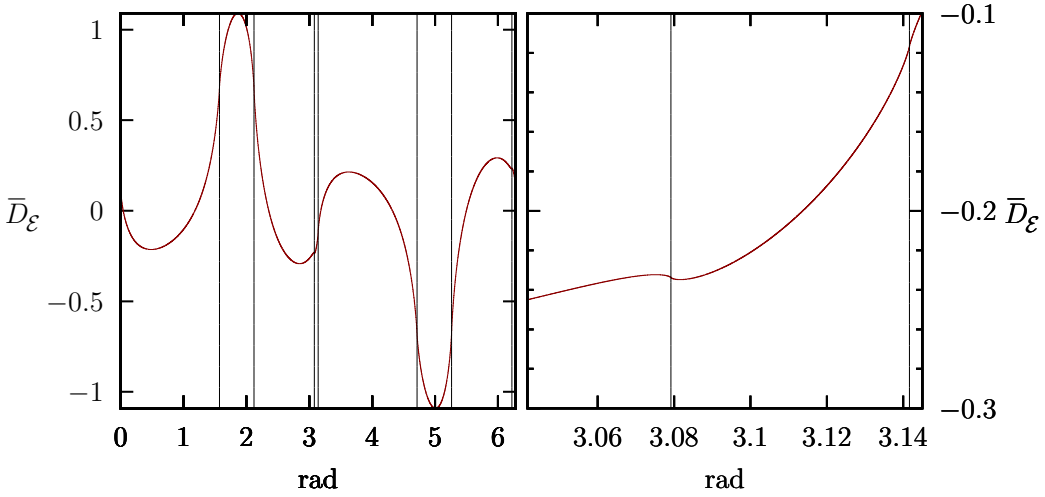

Figure 4.4.: The function $\overline{D}_{\mathcal{E}}$ for $s = \frac{1}{40}$ and $t = \frac{2}{5}$. The diagram on the right shows a zoom on two of the extreme points that are hardly visible on the left. The black vertical bars correspond to the vertices of the black polygon in figure 4.3.

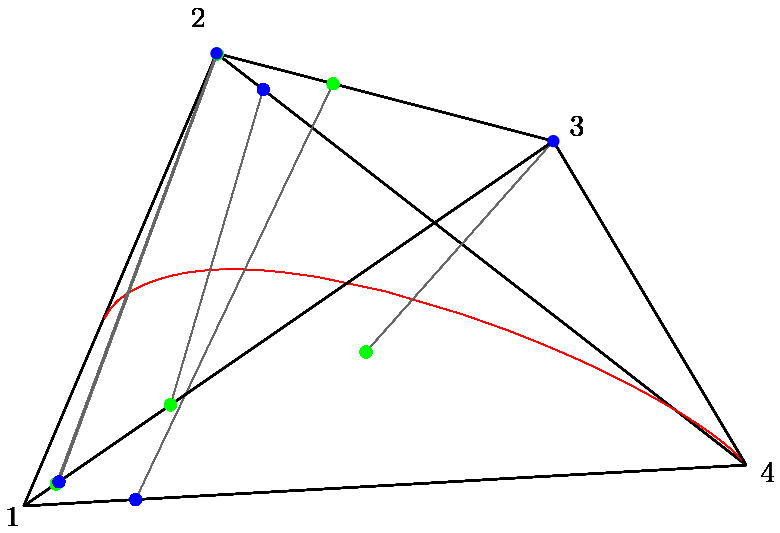

Figure 4.5.: The exponential family (red line) with $s = \frac{1}{40}$ and $t = \frac{2}{5}$. The five local maximizers are marked with blue dots, other projection points are marked with green dots. The rI-projections lie at the intersections of the grey lines with the exponential family.

Theorem 4.6. *Let $\mathcal{E}$ be a one-dimensional exponential family on $\mathcal{X} = \{1, 2, 3, 4\}$, parametrized by a sufficient statistics of the form* (4.7).

(*i*) *If $0 < s < t < 1$, then there is a function $\delta_\nu(s,t)$, defined in* (4.9), *such that the following holds: If $\delta_\nu(s,t) \geq 0$, then the function $\overline{D}_\mathcal{E}$ has six extremal points on $\partial \mathbf{U}_\mathcal{N}$ (three maximizers and three minimizers). Otherwise there are ten extremal points (two additional maximizers and minimizers, respectively).*

(*ii*) *If $0 = s \leq t < 1$ or if $0 < s = t \leq 1$, then there are six extremal points.*

(*iii*) *If $0 = s < t = 1$, then $\overline{D}_\mathcal{E}$ is a linear function, so the set of local maximizers is either all of $\partial \mathbf{U}_\mathcal{N}$ or a proper face of the polytope $\mathbf{U}_\mathcal{N}$. If $\nu = \mathbf{1}$, then $\overline{D}_\mathcal{E} = 0$ on $\mathcal{N}$, so any point on $\partial \mathbf{U}_\mathcal{N}$ is a local maximizer. Conversely, if any element of $\partial \mathbf{U}_\mathcal{N}$ is a local maximizer, then $\mathbf{1}$ is a reference measure of $\mathcal{E}$.*

Figure 4.1 shows a heat map of the function δ_ν for $\nu = \mathbf{1}$. In the chosen parametrization the region with ten extreme values is relatively small. Figures 4.2 and 4.3 show polar plots of the function $\overline{D}_\mathcal{E}$ for two different representative values of s and t. See Example 3.11 for an explanation how to interpret these polar plots.

Figure 4.3 shows the case $s = \frac{1}{40}$ and $t = \frac{2}{5}$. By Figure 4.1 there are ten extreme values of $\overline{D}_\mathcal{E}$. Four of them are hardly visible: They lie very close to one another. This can be seen on Figure 4.4. The vertices of the polygon are marked by vertical lines. They correspond to the points where $\overline{D}_\mathcal{E}$ is not smooth. In fact, at these points the directional derivatives of $\overline{D}_\mathcal{E}$ are infinite, but this is not visible from the diagram.

By Theorem 3.28 the maximizers of $D_\mathcal{E}$ are in bijection with the maximizers of $\overline{D}_\mathcal{E}$. Figure 4.5 shows a typical exponential family with five local maximizers. The three

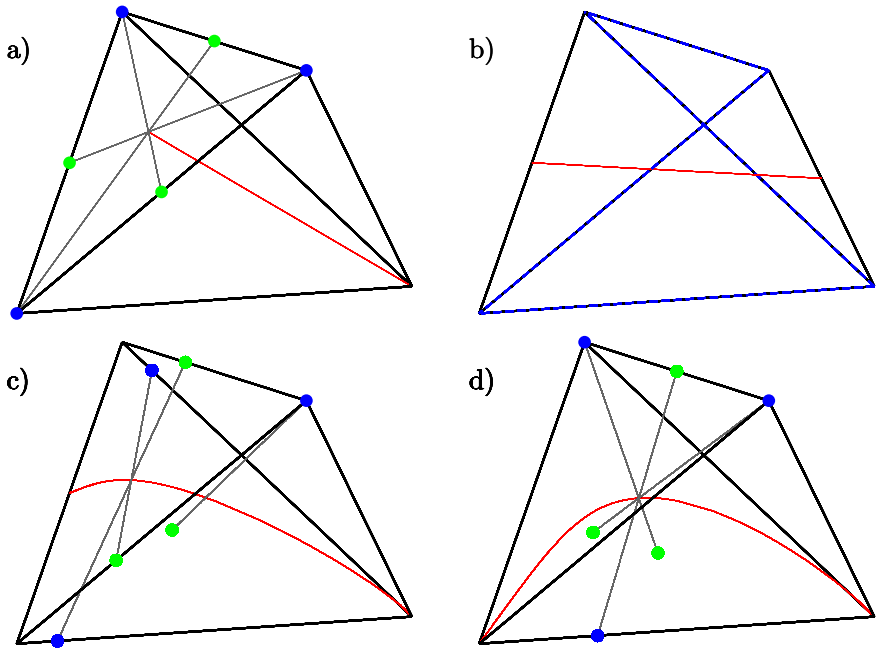

Figure 4.6.: The local maximizers of $D_{\mathcal{E}}$ in the four degenerated cases for $\nu = \mathbf{1}$: a) $0 = s = t < 1$. b) $0 = s < t = 1$. c) $0 = s < t = 0.3 < 1$. d) $0 < s = t = 0.3 < 1$. The labelling and the colours are as in Figure 4.5. In b) the set of maximizers equals the set $\mathrm{K}_{\mathcal{E}}$ of kernel distributions.

local maximizers which are always there are easy to interpret, using Lemma 3.6: Point measures δ_x that are mapped by the moment map into the relative interior of $\mathbf{M}_A$ are always local maxima, by Lemma 3.6. Furthermore, consider the convex hull $\Delta_{1,4}$ of δ_1 and δ_4. By compactness of $\Delta_{1,4}$, there must be a local maximum P in $\Delta_{1,4}^{\circ}$. By Lemma 3.6, P is a local maximum of $D_{\mathcal{E}}$ without constraints.

The degenerated cases are visualized in Figure 4.6. Note that two of the three local maximizers in the case $0 = s < t = 1$ share the same rI-projection, and if $0 < s = t < 1$, then all local maximizers share the same rI-projection. These facts can be easily proved using that two probability measures P, Q have the same rI-projection if and only if $P - Q \in \mathcal{N}$.

Remark 4.7. Theorem 4.6 implies that there are only five possible support sets for the global maximizer. For any of these support sets $\mathcal{Z} \in \{\{2\}, \{3\}, \{1,4\}, \{1,3\}, \{2,4\}\}$ it is possible to find a reference measure ν and parameters $0 \leq s \leq t \leq 1$ such that there is a unique global maximizer of $D_{\mathcal{E}}$, and this maximizer has support $\mathcal{Z}$: In many cases one of the point measure δ_2 and δ_3 will be a global maximizer, see for example Figures 4.2 and 4.3. The binomial model Bin(3), discussed in Section 4.4, is an example of an exponential family where the global maximizer is $\frac{1}{2}(\delta_1 + \delta_4)$, see Figure 4.7 b). To find an example of an exponential family such that the global maximizer is supported on $\{1,3\}$, note that in Figure 4.3 the local maximizer with support $\{1,3\}$ has a larger value of $\overline{D}_{\mathcal{E}}$ than the local maximizer with support $\{3\}$. This motivates to adjust the

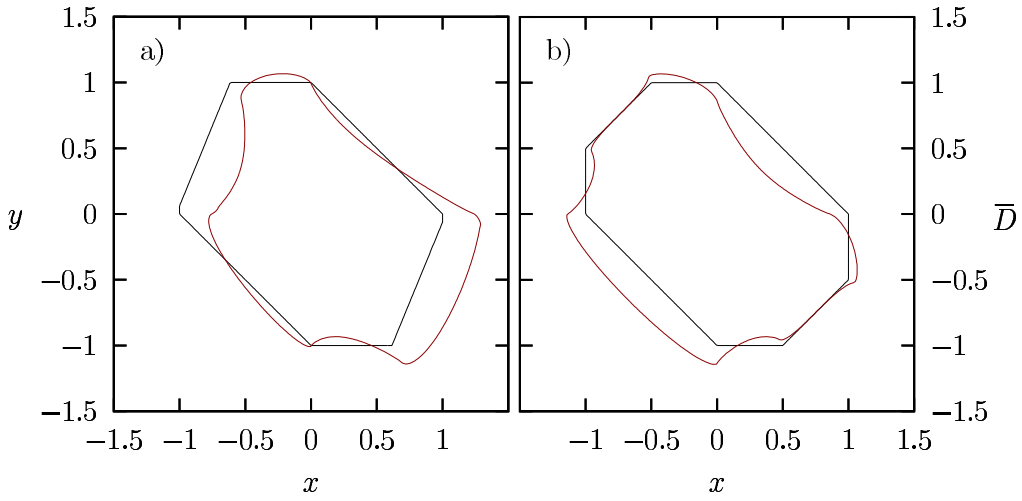

Figure 4.7.: a) A polar plot of $\overline{D}_{\mathcal{E}}$ for $s = \frac{1}{40}$ and $t = \frac{2}{5}$ and $\nu = (2, 4, 1, 2)$. b) A polar plot of $\overline{D}_{\mathcal{E}}$ for Bin 3.

reference measure in such a way to move the exponential family towards $\{2\}$. At the same time it is necessary to make sure that the local maximizer supported on $\{1, 4\}$ is not the global maximizer. A solution is presented in Figure 4.7a).

There do not seem to be one-dimensional exponential families with uniform reference measure such that the global maximizer has support of cardinality two, although this would be possible according to Theorem 4.6 and Proposition 4.4. An empirical search for such exponential families did not yield a positive result.

Proof of Theorem 4.6. The normal space $\mathcal{N}$ is spanned by

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} s-1 & 1 & 0 & -s \\ t-1 & 0 & 1 & -t \end{pmatrix}$$

and can be parametrized by $u(x, y) := xu_1 + yu_2$. The normal space is partitioned by the four hyperplanes

$$x = 0, \qquad y = 0, \qquad sx + ty = 0, \qquad (1-s)x + (1-t)y = 0.$$

If $0 < s < t < 1$, then these four hyperplanes are distinct, and $\partial \mathbf{U}_{\mathcal{N}}$ is an octagon. In this case, there are eight maximal sign vectors. Otherwise, the number of maximal sign vectors is less. If $u \in \partial \mathbf{U}_{\mathcal{N}}$ is a local maximizer, then $-u$ is a local minimizer, and vice versa. Therefore it suffices to find the local maximizers.

First, assume $0 < s < t < 1$. By the results of Section 4.1.2 there are always local maximizers with sign vectors $(+, -, -, +)$, $(-, +, -, -)$ and $(-, -, +, -)$. The interesting sign vectors are $\sigma = (+, -, +, -)$ and its inverse $-\sigma = (-, +, -, +)$. By symmetry it suffices to discuss one of them: Either there are no extremal points of $\overline{D}_{\mathcal{E}}$ with sign vector σ, or there are two extremal points, one local maximizer u_1 and one local minimizer u_2 of $\overline{D}_{\mathcal{E}}$, and in this case $-u_1$ is a local minimizer and $-u_2$ is a local maximizer with sign vector $-\sigma$.

The equations from Section 4.1.2 are applicable: There are two extremal points of $\overline{D}_{\mathcal{E}}$ with sign vector σ if and only if the derivative of $\frac{\partial}{\partial\lambda}\overline{D}_{\mathcal{E}}(u_\lambda)$ takes negative values for some λ. The map $\zeta \mapsto \lambda(\zeta)$ is

$$\lambda(\zeta) = \frac{s\nu_2\zeta^s + \nu_4\zeta}{t(\nu_2\zeta^s + \nu_4\zeta)}.$$

The equation

$$0 = \frac{\partial^2}{\partial\lambda^2}\overline{D}_{\mathcal{E}}(u_\lambda) = C'(\xi)\left(-s(t-s)\nu_2\zeta^s + (1-t)\nu_4\zeta\right)$$

has a unique nonzero solution for $\zeta = \zeta_0 := \left(\frac{\nu_2 s(t-s)}{\nu_4(1-t)}\right)^{\frac{1}{1-s}}$, so

$$\lambda_0 := \lambda(\zeta_0) = \frac{s}{t}\frac{(1-t)\nu_4\zeta_0 + (t-s)\nu_4\zeta_0}{(1-t)\nu_4\zeta_0 + s(t-s)\nu_4\zeta_0} = \frac{s(1-s)}{t(1-t+st-s^2)}.$$

Therefore, the minimum value of $\frac{\partial}{\partial\lambda}\overline{D}_{\mathcal{E}}(u_\lambda)$ equals

$$\begin{aligned}\left.\frac{\partial}{\partial\lambda}\overline{D}_{\mathcal{E}}(u_\lambda)\right|_{\lambda=\lambda_0} &= \log\frac{\nu_1 s(1-s)}{\nu_3((1-s)(t-s)(1-t))\zeta_0^t} \\ &= \frac{1-s-t}{1-s}\log\left(\frac{s}{1-t}\right) - \frac{1-s+t}{1-s}\log(t-s) + \log\left(\left(\frac{\nu_4}{\nu_2}\right)^{\frac{t}{1-s}}\frac{\nu_1}{\nu_3}\right).\end{aligned}$$

There are two critical points of $\overline{D}_{\mathcal{E}}$ in $\partial\mathbf{U}_{\mathcal{N}}$ with sign vector σ if and only if

$$\delta_\nu(s,t) := (1-s-t)\log\left(\frac{s}{1-t}\right) - (1-s+t)\log(t-s) + \log\left(\left(\frac{\nu_4}{\nu_2}\right)^t\left(\frac{\nu_1}{\nu_3}\right)^{1-s}\right) \quad (4.9)$$

is less then zero.

It remains to discuss the degenerated cases. The number of local maximizers follows from the general discussion in Section 4.1.2; in one case there are additional statements to prove:

- If $0 = s < t < 1$, then there are three local maximizers with sign vectors $(+,-,-,+)$, $(-,+,-,+)$ and $(-,-,+,-)$.
- If $0 < s = t < 1$, then there are three local maximizers with sign vectors $(+,-,-,+)$, $(-,+,-,-)$ and $(-,-,+,-)$.
- If $0 = s = t < 1$, then there are three maximizers with sign vectors $(+,-,-,0)$, $(-,+,-,0)$ and $(-,-,+,0)$.
- If $0 = s < t = 1$, then $u(x,y) = (-x,x,y,-y)$. It follows that $\overline{D}_{\mathcal{E}}(u(x,y)) = x\log\frac{\nu_1}{\nu_2} + y\log\frac{\nu_4}{\nu_3}$. The function $\overline{D}_{\mathcal{E}}$ is linear in x and y, and therefore, the set of maximizers is either all of $\partial\mathbf{U}_{\mathcal{N}}$ or a proper face of the polytope $\mathbf{U}_{\mathcal{N}}$. If $\nu_1/\nu_2 = \nu_4/\nu_3 = 1$, then $\overline{D}_{\mathcal{E}} = 0$ is constant, so every point in $\partial\mathbf{U}_{\mathcal{N}}$ is a local maximizer. In this case, let $\xi = \nu_2/\nu_3$. Then $P_\xi = \frac{1}{4}\mathbf{1}$. Therefore, $\nu_1/\nu_2 = \nu_4/\nu_3 = 1$ if and only if $\mathbf{1}$ is a reference measure of $\mathcal{E}$. □

4.1.4. Other low-dimensional exponential families.

The calculations in 4.1.2 generalize to exponential families of higher dimension as follows: First, the possible support sets of a maximizer have to be identified. For any such support set $\mathcal{Z} \subset \mathcal{X}$ choose an affine parametrization $\lambda \mapsto P_\lambda \in \mathbf{P}(\mathcal{Z})$. Let $\mathcal{Y}$ be the smallest facial set containing $\mathcal{Z}$; then $P_\mathcal{E} \in \mathcal{E}^\mathcal{Y} := \overline{\mathcal{E}} \cap \mathbf{P}(\mathcal{Y})^\circ$ for all $P \in \mathbf{P}(\mathcal{Z})$ by Lemma 2.28. Choose a monomial parametrization $\xi \in \mathbb{R}^h \mapsto P_\xi \in \mathcal{E}^\mathcal{Y}$ of $\mathcal{E}^\mathcal{Y}$, see Theorem 2.29. The equation $\pi_A(P_\lambda) = \pi_A(P_\xi)$ is affine in λ and can be solved for $\lambda = \lambda(\xi)$. This solution is unique by Lemma 3.5. The restriction $P_{\lambda(\xi)}(x) > 0$ for all $x \in \mathcal{Z}$ determines a connected region $\Xi \subseteq \mathbb{R}^h$ (in the one-dimensional case, Ξ equals an open interval (ξ_i, ξ_j)). The function $D(P_{\lambda(\xi)} \| P_\xi)$ can be differentiated with respect to ξ or λ and analyzed as above.

By Theorem 3.2 the first order conditions are

$$\log \frac{P_\lambda(i)}{\nu_i \xi^{a_i}} - \log \frac{P_\lambda(j)}{\nu_j \xi^{a_j}}, \qquad \text{for all } i, j \in \mathcal{Z}.$$

These equations are equivalent to

$$P_\lambda(i)\nu_j \xi^{a_j} - P_\lambda(j)\nu_i \xi^{a_i}, \qquad \text{for all } i, j \in \mathcal{Z}. \tag{4.10}$$

In short, the method of this section is to solve the equations $AP_\lambda = AP_\xi$ for $P_\lambda = P_{\lambda(\xi)}$ (this amounts to inverting a submatrix of A) and to plug the solution into equations (4.10). The resulting equations have to be solved for ξ.

This method becomes impractical if the dimension of $\mathcal{E}$ is too large. Let $\dim \mathcal{E} = d$ and $N = |\mathcal{X}|$. Then the dimension of the set $\Xi \subseteq \mathbb{R}^h$ defined above is bounded from above by d, and by Lemma 3.5 the same is true for the dimension of $\mathbf{P}(\mathcal{Z})$. Therefore, essentially a family of d-dimensional problems has to be solved. In contrast, the set $\partial \mathbf{U}_\mathcal{N}$ is $(N - d - 2)$-dimensional. Hence, if the codimension of $\mathcal{E}$ is low, then the algorithm from Section 3.6 is preferable.

The function $\overline{D}_\mathcal{E}$ can be treated similarly: Let σ be a critical sign vector, and let $\mathcal{Z} = \operatorname{supp}(\sigma^+)$ and $\mathcal{Y} = \operatorname{supp}(\sigma)$. Define a parametrization $\lambda \mapsto u^+_\lambda \in \mathbf{P}(\mathcal{Z})$, and let $\zeta \mapsto u^-_\zeta$ be a monomial parametrization of the exponential family $\mathcal{E}^{\mathcal{Y}\setminus\mathcal{Z}} = \{Q^{\mathcal{Y}\setminus\mathcal{Z}} : Q \in \mathcal{E}\}$. The linear equation $Au^+_\lambda = Au^-_\zeta$ can be solved for $u^+_\lambda = u^+_{\lambda(\zeta)}$. The first order conditions of $\overline{D}_\mathcal{E}(u_{\lambda(\zeta)})$ are equivalent to

$$u^+_\lambda(i)\nu_j \zeta^{a_j} - u^+_\lambda(j)\nu_i \zeta^{a_i}, \qquad \text{for all } i, j \in \mathcal{Z}. \tag{4.11}$$

Then one can plug $u^+_{\lambda(\zeta)}$ into (4.11) and solve the resulting equations for ζ. This method also performs better if the dimension of the exponential family $\mathcal{E}$ is small, but it has a slight advantage over the other method: If $\mathcal{Y} \setminus \mathcal{X} = \operatorname{supp}(u^-_\zeta)$ is small, then it may happen that the dimension of $\mathcal{E}^{\mathcal{Y}\setminus\mathcal{Z}}$ is less than $d := \dim \mathcal{E}$, and for such critical sign vectors σ the dimension of the problem is less than d.

As in Sections 3.6 and 3.7 the corresponding equations for both methods proposed in this section can be reformulated as algebraic equations if the exponential family $\mathcal{E}$

is algebraic. As in Section 3.7, a nonnegative integer sufficient statistics A has to be chosen, and then the maps $\xi \mapsto P_\xi$ and $\zeta \mapsto u_\zeta^-$ are monomial.

In principle, the arguments from the proof of Proposition 4.4 concerning the global maximizers also generalize to higher dimensions: Comparing different probability measures with the same rI-projection gives restrictions on the possible support sets (a related idea will be used in the proof of Theorem 5.5). It is not easy to make general statements, though. The discussion of the one-dimensional case is simplified by the fact that one-dimensional point configurations are easy to parametrize, and the level sets of the entropy of a binary random variable are easy to understand.

4.2. Partition models

The partition exponential families introduced in this section are a class of exponential families, and their closures are called partition models. They arise naturally in the study of symmetries, and they approximate arbitrary probability distributions in an optimal way, in a sense that will be explained in Section 5.2. Partition exponential families are convex exponential families, and the information divergence from convex exponential families has been studied in [52]. Apart from this, partition exponential families do not seem to have been studied before, despite their peculiar properties. In other contexts the name "partition model" is used for other mathematical objects, but there seems to be little danger of confusion.

Definition 4.8. A *partition* $\mathcal{X}'$ of $\mathcal{X}$ is a set $\mathcal{X}' = \{\mathcal{X}^1, \mathcal{X}^2, \ldots, \mathcal{X}^{N'}\}$ of nonempty subsets $\mathcal{X}^i \subset \mathcal{X}$ such that $\mathcal{X} = \mathcal{X}^1 \cup \mathcal{X}^2 \cup \cdots \cup \mathcal{X}^{N'}$ and $\mathcal{X}^i \cap \mathcal{X}^j = \emptyset$ for all $1 \le i < j \le N'$. The subsets $\mathcal{X}^i \subseteq \mathcal{X}$ are called the *blocks* of the partition $\mathcal{X}'$. For any $x \in \mathcal{X}$ the block $\mathcal{X}^i$ containing x is denoted by $\mathcal{X}^x$.

The *coarseness* $c(\mathcal{X}')$ of a partition $\mathcal{X}'$ is the cardinality of the largest block of $\mathcal{X}'$. A partition $\mathcal{X}'$ is called *homogeneous* if all blocks of $\mathcal{X}'$ have the same cardinality $c(\mathcal{X}')$. Partitions are in bijection with equivalence relations, the blocks of a partition corresponding to the equivalence classes. The equivalence relation corresponding to the partition $\mathcal{X}'$ is denoted by $\sim_{\mathcal{X}'}$.

Let $\mathcal{X}'$ be a partition of $\mathcal{X}$, and let $A^{\mathcal{X}'} \subset \mathbb{R}^{N' \times N}$ be the matrix with entries

$$A^{\mathcal{X}'}_{i,x} = \begin{cases} 1, & \text{if } x \in \mathcal{X}^i, \\ 0, & \text{else.} \end{cases}$$

This matrix has the following interpretation: For any probability measure P on $\mathcal{X}$ let P' be the probability measure on $\mathcal{X}'$ induced by P. This means that P' satisfies $P'(\mathcal{X}^i) = P(\mathcal{X}^i)$ for all $i = 1, \ldots, N'$. Then $P' = A^{\mathcal{X}'} P$.

Definition 4.9. Let $\mathcal{X}'$ be a partition of $\mathcal{X}$. The exponential family $\mathcal{E}_{\mathcal{X}'} := \mathcal{E}_{\mathbf{1}, A^{\mathcal{X}'}}$ with reference measure $\mathbf{1}$ and sufficient statistics $A^{\mathcal{X}'}$ is called the *partition exponential family* of $\mathcal{X}'$, and $\overline{\mathcal{E}_{\mathcal{X}'}}$ is the *partition model* of $\mathcal{X}'$.

Partition models are, in fact, also linear families: $\overline{\mathcal{E}_{\mathcal{X}'}}$ consists of all probability measures P that satisfy $P(x) = P(y)$ whenever $x \sim_{\mathcal{X}'} y$. In particular, partition exponential families are convex exponential families. Convex exponential families and their maximizers have been studied by Ay and Matúš in [52], which contains more detailed arguments for the following calculations. It follows from [52, Proposition 1] that a convex exponential family is a partition exponential family if and only if it contains the uniform distribution.

The convex support of a partition exponential family is a simplex of dimension $N' - 1$ with vertex set $\{A_x^{\mathcal{X}'} : x \in \mathcal{X}\}$, since each vector $A_x^{\mathcal{X}'}$ is actually a unit vector in $\mathbb{R}^{N'}$. The converse is also true:

Lemma 4.10. *An exponential family with uniform reference measure and sufficient statistics $A \in \mathbb{R}^{h \times \mathcal{X}}$ is a partition exponential family if and only if its convex support is a simplex with vertex set $\{A_x : x \in \mathcal{X}\}$.*

Proof. Assume that the convex support $\mathbf{M}_A$ of $\mathcal{E}_{\mathbf{1},A}$ is a simplex with vertex set $\{A_x : x \in \mathcal{X}\}$. Define an equivalence relation $\sim$ on $\mathcal{X}$ via $x \sim y$ if and only if $A_x = A_y$. Then $\mathcal{E}_{\mathbf{1},A}$ equals the partition exponential family of $\sim$ by the last statement of Lemma 2.8. □

Remark 4.11. Composite systems have natural homogeneous partitions, which lead to hierarchical models (see Section 2.4 for the notation): Suppose that $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ and let $K \subseteq \{1, \ldots, n\}$. Then K induces an equivalence $\sim_K$ on $\mathcal{X}$ via $x \sim_K y$ if and only if $x_i = y_i$ for all $i \in K$. The equivalence classes of $\sim_K$ form a homogeneous partition $\mathcal{X}^K$ of $\mathcal{X}$ of coarseness $\prod_{i:i \notin K} N_i$. The corresponding partition model $\overline{\mathcal{E}_K}$ consists of those probability distributions P satisfying $P(x) = P(y)$ whenever $x \sim_K y$. Therefore, $\mathcal{E}_K$ equals the hierarchical exponential family $\mathcal{E}_{\{K\}}$. Conversely, any homogeneous partition $\mathcal{X}'$ can be used to find a bijection of $\mathcal{X}$ with a composite system $\mathcal{X}_1 \times \mathcal{X}_2$, where $\mathcal{X}_1 = \mathcal{X}'$ and $\mathcal{X}_2 \in \mathcal{X}'$. Then the partition $\mathcal{X}'$ arises from $\sim_K$, where $K = \{1\}$.

Remark 4.12. Partition models can be used to model symmetries. This was first noted by Juríček, who used this idea to compute the global maximizers of $D_{\mathcal{E}}$ for the multinomial models [41]. If a symmetry group G acts on $\mathcal{X}$, then it induces a partition $\mathcal{X}^G$ of $\mathcal{X}$ into orbits $\mathcal{X}^1, \ldots, \mathcal{X}^{N'}$. The action of G extends naturally to an action on $\mathbb{R}^{\mathcal{X}}$. Any exponential family that consists of G-invariant probability measures is a subfamily of $\mathcal{E}_{\mathcal{X}^G}$ (such exponential families are called *G-exchangeable* in [41]). Conversely, an arbitrary partition model $\overline{\mathcal{E}_{\mathcal{X}'}}$ arises in this way from the group of all permutations g of $\mathcal{X}$ such that $g(\mathcal{X}^i) = \mathcal{X}^i$ for all $\mathcal{X}^i \in \mathcal{X}'$.

Remark 4.13. The natural map $\phi : x \mapsto \mathcal{X}^x$ induces a pushforward $\phi_* : \mathbf{P}(\mathcal{X}) \mapsto \mathbf{P}(\mathcal{X}')$ via $\phi_* P(\mathcal{X}^j) = \sum_{y \in \mathcal{X}^j} P(y)$. This map agrees with the moment map $\pi_{A^{\mathcal{X}'}}$ and restricts to a bijection of the partition model $\overline{\mathcal{E}_{\mathcal{X}'}}$ with the probability simplex $\mathbf{P}(\mathcal{X}')$. This bijection preserves the information divergence: If $P, Q \in \overline{\mathcal{E}_{\mathcal{X}'}}$, then $P(x)/Q(x) =$

$P(\mathcal{X}^x)/Q(\mathcal{X}^x)$ for all $x \in \mathcal{X}$, whence

$$D(\phi_* P \| \phi_* Q) = \sum_{j=1}^{N'} P(\mathcal{X}^j) \log \frac{P(\mathcal{X}^j)}{Q(\mathcal{X}^j)} = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} = D(P\|Q).$$

This invariance property was first noted by Juríček when he studied symmetries [41]. Under this bijection ϕ_*, exponential families on $\mathcal{X}'$ are in bijection with exponential subfamilies of $\mathcal{E}_{\mathcal{X}'}$: If $\mathcal{E}$ is an exponential subfamily of $\mathcal{E}_{\mathcal{X}'}$ with reference measure ν and sufficient statistics $A \in \mathbb{R}^{h \times \mathcal{X}}$, then $\nu_x = \nu_y$ and $A_x = A_y$ whenever $x \sim_{\mathcal{X}'} y$. The bijection ϕ_* maps $\mathcal{E}$ to the exponential family with sufficient statistics $A' \in \mathbb{R}^{h \times \mathcal{X}'}$ given by

$$(A')_{i,\mathcal{X}^j} = A_{i,x}, \text{ for all } x \in \mathcal{X}^j,$$

and reference measure $\nu_G = \phi_* \nu$. The uniform reference measure is mapped to an integer reference measure. This is another motivation for allowing arbitrary reference measures. Since the map ϕ_* preserves the information divergence, it can be used to relate the problem of maximizing the information divergence $D_{\mathcal{E}}(Q)$ from an exponential subfamily $\mathcal{E} \subseteq \mathcal{E}_{\mathcal{X}'}$ subject to the constraint that $Q \in \overline{\mathcal{E}_{\mathcal{X}'}}$ with the problem of maximizing $D_{\phi_*(\mathcal{E})}$ over $\mathbf{P}(\mathcal{X}')$.

An example to the last three remarks is given by the binary i.i.d. models and the binomial models, which will be studied in Section 4.4.

For partition models it is possible to find the mapping $P \mapsto P_{\mathcal{E}}$ explicitly: The rI-projection $P_{\mathcal{E}}$ of $P \in \mathbf{P}(\mathcal{X})$ satisfies $A^{\mathcal{X}'} P_{\mathcal{E}} = A^{\mathcal{X}'} P$. This implies $P(\mathcal{X}^i) = P_{\mathcal{E}}(\mathcal{X}^i)$ for $i = 1, \ldots, N'$. Therefore,

$$P_{\mathcal{E}}(x) = P_{\mathcal{E}}(x|\mathcal{X}^x) P(\mathcal{X}^x), \qquad \text{for all } x \in \mathcal{X}. \tag{4.12}$$

Since $P_{\mathcal{E}}$ maximizes the entropy subject to (4.12), it follows that $P_{\mathcal{E}}(\cdot|\mathcal{X}^x) = \frac{1}{|\mathcal{X}^x|} \mathbf{1}_{\mathcal{X}^x}$ is the uniform distribution on $\mathcal{X}^x$. If the partition comes from some symmetry group G, as in Remark 4.12, then the rI-projection map $P \mapsto P_{\mathcal{E}}$ is the symmetrization map with respect to the action of G.

From (4.12) and $P(x) = P(x|\mathcal{X}^x) P(\mathcal{X}^x)$ it follows that

$$D_{\mathcal{E}}(P) = \sum_{i=1}^{N'} P(\mathcal{X}^i) D(P(\cdot|\mathcal{X}^i) \| \frac{1}{|\mathcal{X}^x|} \mathbf{1}_{\mathcal{X}^x}) = \sum_{i=1}^{N'} P(\mathcal{X}^i) \left(\log |\mathcal{X}^x| - H(P(\cdot|\mathcal{X}^i)) \right). \tag{4.13}$$

As a consequence:

Lemma 4.14. *If $\overline{\mathcal{E}}$ is a partition model of a partition $\mathcal{X}^1, \ldots, \mathcal{X}^{N'}$ of coarseness c, then $\max D_{\mathcal{E}} = \log(c)$. A probability measure $P \in \mathbf{P}(\mathcal{X})$ maximizes $D_{\mathcal{E}}$ if and only if the following two conditions are satisfied:*

(i) $P(\mathcal{X}^i) > 0$ only if $|\mathcal{X}^i| = c$.

(ii) $P(\cdot|\mathcal{X}^i)$ is a point measure for all i such that $|\mathcal{X}^i| = c$.

Corollary 4.15. *Let $\overline{\mathcal{E}}$ be the partition model of a partition $\mathcal{X}'$ of coarseness c, and let $\mathcal{Z}$ be the union of the blocks of $\mathcal{X}'$ of cardinality c. Then any $Q \in \overline{\mathcal{E}}$ with support contained in $\mathcal{Z}$ is the rI-projection of some global maximizer of $D_{\mathcal{E}}$. In particular, if $\mathcal{X}'$ is homogeneous, then any $Q \in \overline{\mathcal{E}}$ is the rI-projection of some global maximizer of $D_{\mathcal{E}}$.*

Proof. For any $\mathcal{X}^i \in \mathcal{X}'$ of cardinality c choose a representative $x_i \in \mathcal{X}^i$. Define $P \in \mathbf{P}(\mathcal{X})$ by $P(\mathcal{X}^i) = Q(\mathcal{X}^i)$ and $P(\cdot|\mathcal{X}^i) = \delta_{x_i}$ for all i such that $|\mathcal{X}^i| = c$. Then $P_{\mathcal{E}} = Q$, so the statement follows from Lemma 4.14. □

Let $\mathcal{E}$ be an exponential subfamily of a partition model $\mathcal{E}_{\mathcal{X}'}$. For $P \in \mathbf{P}(\mathcal{X})$ denote by $P_{\mathcal{X}'}$ the rI-projection of P. By the Pythagorean identity, $D_{\mathcal{E}}(P) = D(P\|P_{\mathcal{X}'}) + D_{\mathcal{E}}(P_{\mathcal{X}'})$. If P is a local maximizer of $D_{\mathcal{E}}$, then $P(\cdot|\mathcal{X}^j)$ is a point measure for all j by Lemma 3.5 and Remark 4.13. Therefore, (4.13) implies

$$D_{\mathcal{E}}(P) = \sum_{i=1}^{N'} P(\mathcal{X}^i) \log |\mathcal{X}^x| + D_{\mathcal{E}}(P_{\mathcal{X}'}). \tag{4.14}$$

In this sense, the maximization of $D(P\|\mathcal{E})$ for $P \in \mathbf{P}(\mathcal{X})$ differs from the maximization of $D(P\|\mathcal{E})$ for $P \in \mathcal{E}_{\mathcal{X}'}$ just by a piece-wise linear function. See Section 4.4 for an example. If the partition $\mathcal{X}'$ is homogeneous, then the two maximization problems are equivalent in the sense that the solution of one yields a solution of the other:

Lemma 4.16. *Let $\mathcal{X}'$ be a homogeneous partition of $\mathcal{X}$ of coarseness c, and let $\mathcal{E}$ be an exponential subfamily of the partition model $\mathcal{E}_{\mathcal{X}'}$. Then*

$$\max D_{\mathcal{E}} = \log(c) + \max\{D_{\mathcal{E}}(Q) : Q \in \mathcal{E}_{\mathcal{X}'}\}.$$

Proof. Choose $Q \in \mathcal{E}_{\mathcal{X}'}$ such that $D_{\mathcal{E}}(Q)$ is maximal. By Corollary 4.15 there exists $P \in \mathbf{P}(\mathcal{X})$ such that $Q = P_{\mathcal{X}'}$ and $D(P\|Q) = \log(c)$. Then $D_{\mathcal{E}}(P) = \log(c) + \max\{D_{\mathcal{E}}(Q) : Q \in \mathcal{E}_{\mathcal{X}'}\}$. Conversely, let $P \in \mathbf{P}(\mathcal{X})$. By Lemma 4.14,

$$D_{\mathcal{E}}(P) = D(P\|P_{\mathcal{X}'}) + D_{\mathcal{E}}(P_{\mathcal{X}'}) \leq \log(c) + \max\{D_{\mathcal{E}}(Q) : Q \in \mathcal{E}_{\mathcal{X}'}\},$$

proving the converse inequality. □

The following lemma, which is due to Matúš [51, Remark 1], is an application of Lemma 4.16 and Remark 4.13 to hierarchical models.

Lemma 4.17. *Let $\mathcal{X} = \prod_{i=1}^n \mathcal{X}_i$, let Δ be a collection of subsets of $[n]$, and let $K = \cup_{I\in\Delta} I$. Denote by $\mathcal{E}_{K,\Delta}$ the hierarchical exponential family on $\mathcal{X}_K$ specified by Δ. Then*

$$\max_{Q\in\mathbf{P}(\mathcal{X})} D(Q\|\mathcal{E}_\Delta) = \sum_{i\notin K} \log|\mathcal{X}_i| + \max_{Q\in\mathbf{P}(\mathcal{X}_K)} D(Q\|\mathcal{E}_{K,\Delta}).$$

4.3. Exponential families with $\max D_{\mathcal{E}} = \log(2)$

By Remark 3.15 the value of $\max D_{\mathcal{E}}$ is at least $\log(2)$ for all exponential families $\mathcal{E} \subsetneq \mathbf{P}(\mathcal{X})^\circ$. This section studies exponential families $\mathcal{E}$ where $\max D_{\mathcal{E}} = \log(2)$. For such an exponential family, any kernel measure is a local maximizer of $D_{\mathcal{E}}$. Furthermore, $\overline{D}_{\mathcal{E}}(u) = 0$ for all $u \in \mathcal{N}$ (even if $u \notin \mathbf{U}_{\mathcal{N}}$). The main results are:

Theorem 4.18. *Let $\mathcal{E}$ be an exponential family on a finite set $\mathcal{X}$ of cardinality N. If $\max D_{\mathcal{E}} = \log(2)$, then the dimension of $\mathcal{E}$ is at least $\lceil \frac{N}{2} \rceil - 1$.*

Theorem 4.19. *Let $\mathcal{X}$ be a finite set of cardinality N, and let $\mathcal{E}$ be an exponential family on $\mathcal{X}$ of dimension $\lceil \frac{N}{2} \rceil - 1$ satisfying $\max D_{\mathcal{E}} = \log(2)$. If N is even, then $\mathcal{E}$ is a partition model. If N is odd, then there is a set $\mathcal{Z} \subseteq \mathcal{X}$ of cardinality three, a partition model $\overline{\mathcal{E}_{\mathcal{X}\setminus\mathcal{Z}}}$ on $\mathcal{X}\setminus\mathcal{Z}$ and a one-dimensional exponential family $\mathcal{E}_{\mathcal{Z}}$ on $\mathcal{Z}$ such that $\max D(\cdot\|\mathcal{E}_{\mathcal{X}\setminus\mathcal{Z}}) = \log(2) = \max D(\cdot\|\mathcal{E}_{\mathcal{Z}})$, and the closure $\overline{\mathcal{E}}$ equals the mixture of $\overline{\mathcal{E}_{\mathcal{X}\setminus\mathcal{Z}}}$ and $\overline{\mathcal{E}_{\mathcal{Z}}}$. If $\mathbf{1}$ is a reference measure of $\mathcal{E}$, then $\overline{\mathcal{E}}$ is a partition model.*

Proposition 4.20. *Let $\mathcal{X} = \{1, 2, 3\}$. For any $u \in \mathbb{R}^{\mathcal{X}}$ such that $u_1 + u_2 + u_3 = 0$ there exists a unique exponential family $\mathcal{E}$ on $\mathcal{X}$ with normal space $\mathcal{N} = \mathbb{R}u$ such that $\max D_{\mathcal{E}} = \log(2)$.*

The proofs of the three results will be given below after a series of preliminary lemmas. Under the additional assumption that N is even Theorem 4.19 has a much simpler proof, see Theorem 5.5.

Let $\mathcal{E}$ be an exponential family with sufficient statistics A and normal space $\mathcal{N}$.

Lemma 4.21. *For any $v_0, v_1, \ldots, v_s \in \mathcal{N}$ let $\mathcal{Z} = \operatorname{supp}(v_0) \setminus \cup_{j=1}^{s} \operatorname{supp}(v_j)$. Suppose that $\max D_{\mathcal{E}} = \log(2)$. Then*

$$\sum_{x\in\mathcal{Z}} v(x) \log \frac{|v(x)|}{\nu_x} = 0 \quad \text{and} \quad \sum_{x\in\mathcal{Z}} v(x) = 0 \quad \text{for all } v \in \mathcal{N}.$$

Proof. The proof is by induction on s. Let $s = 0$. Any $v_0 \in \mathcal{N}$ is a critical point of $\overline{D}_{\mathcal{E}}$. The equality $v(\mathcal{Z}) = 0$ for all $v \in \mathcal{N}$ follows from Proposition 3.21 (i). Let $\mathcal{Z}' = \mathcal{X} \setminus \mathcal{Z}$. Proposition 3.21 (iii) implies that

$$\sum_{x\in\mathcal{Z}'} v(x) \log \frac{|v(x)|}{\nu_x} \le v^+(\mathcal{Z}')\overline{D}_{\mathcal{E}}(v_0) = 0 \qquad \text{for all } v \in \mathcal{N}.$$

In this inequality v may be replaced by $-v$, showing that $\sum_{x\in\mathcal{Z}'} v(x) \log \frac{|v(x)|}{\nu_x} = 0$. Hence $\sum_{x\in\mathcal{Z}} v(x) \log \frac{|v(x)|}{\nu_x} = \overline{D}_{\mathcal{E}}(v) - \sum_{x\in\mathcal{Z}'} v(x) \log \frac{|v(x)|}{\nu_x} = 0$.

If $s \ge 1$, then let $\mathcal{Y} = \mathcal{X} \setminus \operatorname{supp}(v_s)$. Since v_s is a critical point of $\overline{D}_{\mathcal{E}}$, the set $\mathcal{Y}$ is facial. By Lemma 3.4 the set $\mathcal{N}' = \{v|_{\mathcal{Y}} : v \in \mathcal{N}\}$ is the normal space of an exponential family $\mathcal{E}'$, and the case $s = 0$ implies $\overline{D}_{\mathcal{E}'}(w) = \overline{D}_{\mathcal{E}}(w) = 0$ for all $w \in \mathcal{N}'$. Therefore, the statement follows from induction. □

Let $\mathcal{X}' = \{x \in \mathcal{X} : v(x) \neq 0 \text{ for some } v \in \mathcal{N}\}$. Define an equivalence relation $\sim$ on $\mathcal{X}'$ via

$$x \sim y \quad \iff \quad v(y) \neq 0 \text{ for all } v \in \mathcal{N} \text{ such that } v(x) \neq 0.$$

It is easy to see that this relation is indeed an equivalence relation: If there exists $v, w \in \mathcal{N}$ such that $v(y) \neq 0 = v(x)$ and $w(x) \neq 0 \neq w(y)$, then $u := v(y)w - w(y)v \in \mathcal{N}$ satisfies $u(y) = 0 \neq u(x)$. In the language of matroid theory (see Appendix A.2) the equivalence classes are the coparallel classes.

Lemma 4.22. *A subset $\mathcal{Z} \subseteq \mathcal{X}$ is an equivalence class of $\sim$ if and only if there exist circuits $\sigma_0, \sigma_1, \ldots, \sigma_s$ of $\mathcal{N}$ such that*

$$\mathcal{Z} = \sigma_0 \setminus \cup_{j=1}^{s} \sigma_j,$$

and such that $\mathcal{Z} \setminus \sigma \in \{\emptyset, \mathcal{Z}\}$ for all circuits σ of $\mathcal{N}$.

Proof. If $x \not\sim y$ for some $y \in \mathcal{X}$, then there exists a $v \in \mathcal{N}$ such that $v(x) \neq 0$ and $v(y) = 0$. By Lemma A.5 there exists a circuit with the same property. Conversely, if $y \sim x$, then $y \in \sigma$ for any circuit σ such that $x \in \sigma$. □

Let $C \in \mathbb{R}^{c \times \mathcal{X}}$ be a matrix such that the rows $c_1, \ldots, c_c$ of C form a circuit basis of $\mathcal{N}$. Since each circuit basis contains a basis, the rank of C equals the dimension of $\mathcal{N}$. The columns of C are denoted by $\{C_x\}_{x \in \mathcal{X}}$.

Lemma 4.23. *Let $\mathcal{Z}$ be an equivalence class of $\sim$. The rank of the submatrix $C|_{\mathcal{Z}}$ consisting of those columns C_x indexed by $\mathcal{Z}$ is one.*

Proof. Let $\mathcal{Z} \subseteq \mathcal{X}$. If the rank of $C|_{\mathcal{Z}}$ is larger than one, then there exist two circuit vectors c_1, c_2 such that $c_1|_{\mathcal{Z}}$ and $c_2|_{\mathcal{Z}}$ are linearly independent and have support $\mathcal{Z}$. Let $x \in \mathcal{Z}$. Let $v = c_2(x)c_1 - c_1(x)c_2 \in \mathcal{N}$. Then $v|_{\mathcal{Z}} \neq 0$ and $\operatorname{supp}(v|_{\mathcal{Z}}) \subseteq \mathcal{Z} \setminus \{x\}$. Therefore, $\mathcal{Z}$ is not an equivalence class of $\sim$. □

The main argument of the last proof can be reformulated in terms of the weak elimination axiom of matroid theory, cf. Appendix A.2. In the language of matroid theory Lemma 4.23 states that the coparallel classes of a matroid have corank one.

Proof of Theorem 4.18. Suppose $\max D_{\mathcal{E}} = \log(2)$. By Lemma 4.23, the rank of C is bounded from above by the number of equivalence classes of $\sim$. Let $\mathcal{Z}$ be an equivalence class of $\sim$. By definition, the submatrix $C|_{\mathcal{Z}} \in \mathbb{R}^{c \times \mathcal{Z}}$ is not the zero matrix. By Lemmas 4.21 and 4.22 the rows $c_i|_{\mathcal{Z}}$ of $C|_{\mathcal{Z}}$ satisfy $\sum_{x \in \mathcal{Z}} c_i(x) = 0$. Hence each equivalence class must contain at least two elements. Therefore, the rank of C, which equals the codimension of $\mathcal{E}$, is bounded from above by $\lfloor \frac{N}{2} \rfloor$, and so the dimension of $\mathcal{E}$ is bounded from below by $N - 1 - \lfloor \frac{N}{2} \rfloor = \lceil \frac{N}{2} \rceil - 1$. □

Lemma 4.24. *If the dimension of $\mathcal{N}$ equals the number of equivalence classes of $\sim$, then the equivalence classes are the circuits of $\mathcal{N}$. In other words, the circuit vectors $c_1, \ldots, c_c$ of a circuit basis are in bijection with the equivalence classes $\mathcal{Z}_1, \ldots, \mathcal{Z}_c$, such that $\mathcal{Z}_i = \operatorname{supp}(c_i)$. Hence $\overline{\mathcal{E}}$ is the mixture of $\overline{\mathcal{E}}_1, \ldots, \overline{\mathcal{E}_c}$, where $\mathcal{E}_c$ is the exponential family $\overline{\mathcal{E}} \cap \mathbf{P}(\mathcal{Z}_i)^\circ$.*

Proof. Let $\mathcal{Z}_1, \ldots, \mathcal{Z}_{c'}$ be the set of equivalence classes of $\sim$. Reorder $\mathcal{X}$ such that the equivalence classes are given by consecutive numbers. Let $\tilde{C}$ be the matrix obtained from C by doing a Gauss elimination through row operations, such that $\tilde{C}$ has c' nonzero rows. By Lemma 4.23, the ith row $\tilde{c}_i$ of $\tilde{C}$ has support contained in $\mathcal{Z}_i \cup \cdots \cup \mathcal{Z}_{c'}$. In particular, $\operatorname{supp}(\tilde{c}_{c'}) = \mathcal{Z}_{c'}$. Therefore, $\tilde{c}_{c'}$ is a circuit vector. If $v \in \mathcal{N}$ has $v(x) \neq 0$ for some $x \in \mathcal{Z}_{c'}$, then $\tilde{v} = v - \frac{v(x)}{\tilde{c}_{c'}(x)}\tilde{c}_{c'}(x)$ satisfies $\operatorname{supp}(\tilde{v}) = \operatorname{supp}(v) \setminus \mathcal{Z}_{c'}$. Hence no other circuit intersects $\mathcal{Z}_{c'}$. By induction, $\operatorname{supp}(c_i)$ equals an equivalence class of $\sim$ for each i. The first statement follows from $\operatorname{supp}(c_i) \neq \operatorname{supp}(c_j)$ for $1 \leq i < j \leq c$. The last statement is a consequence of Corollary 2.27. □

Proof of Theorem 4.19. Assume that the dimension of $\mathcal{E}$ equals $\lceil \frac{N}{2} \rceil - 1$. By the proof of Theorem 4.18 there must be $m := \lfloor \frac{N}{2} \rfloor$ equivalence classes of $\sim$. If N is even, then each equivalence class has cardinality two. If N is odd, then there may be one equivalence class $\mathcal{Z}$ of cardinality three. In this case, reorder $\mathcal{X}$ such that $\mathcal{Z} = \{N-2, N-1, N\}$. By Lemma 4.24 there exists a circuit vector $c \in \mathcal{N}$ such that $\operatorname{supp}(c) = \mathcal{Z}$. Assume without loss of generality that c_{N-2} and c_{N-1} are positive and that $c_N = -(c_{N-1} + c_{N-2}) = -1$. Then

$$\sum_{i=N-2}^{N} c_i \log |c_i| = -h(c_{N-1}, c_{N-2}) \neq 0,$$

where $h(p, q)$ is the entropy of a binary random variable with probabilities p, q. Therefore, if N is even or if $\mathbf{1}$ is a reference measure of $\mathcal{E}$, then all equivalence classes of $\sim$ have cardinality two.

By Lemma 4.24 there are exponential families $\mathcal{E}_1, \ldots, \mathcal{E}_c$ such that $\mathcal{E}_i \subseteq \mathbf{P}(\mathcal{Z}_i)^\circ$ for $i = 1, \ldots, c$ and such that $\overline{\mathcal{E}}$ is the mixture of $\overline{\mathcal{E}_1}, \ldots, \overline{\mathcal{E}_c}$. For $i = 1, \ldots, c$ there is a unique circuit vector with support $\mathcal{Z}_i$, hence $\mathcal{E}_i \neq \mathbf{P}(\mathcal{Z}_i)^\circ$, so $\mathcal{E}_i$ has dimension $|\mathcal{Z}_i| - 1$. If $|\mathcal{Z}_i| = 2$, then $\mathcal{E}_i$ consists of the uniform distribution $\frac{1}{2}\mathbf{1}_{\mathcal{Z}_i}$ on $\mathcal{Z}_i$, so $\overline{\mathcal{E}_i}$ is a partition model, and also the mixture of $\overline{\mathcal{E}_i}$ for those i satisfying $|\mathcal{Z}_i| = 2$ is a partition model. □

Proof of Proposition 4.20. Let $\mathcal{E}$ be a one-dimensional exponential family with normal space $\mathbb{R}u$. By Example 3.13 the set of local maximizers of $D_{\mathcal{E}}$ consists of u^+ and u^-, and they are both projection points. $\mathcal{E}$ satisfies $\max D_{\mathcal{E}} = \log(2)$ if and only if $(u^+)_{\mathcal{E}} = (u^-)_{\mathcal{E}} = \frac{1}{2}(u^+ + u^-)$, which happens if and only if $u^+ + u^-$ is a reference measure of $\mathcal{E}$, proving existence and uniqueness. □

4.4. Binary i.i.d. models and binomial models

This section discusses the binary i.i.d. families and the binomial families, two related classes of one-dimensional exponential families. The goal of this section is threefold: First, it illustrates how partition models arise in the study of symmetries, see Remarks 4.12 and 4.13. Second, it shows how to extend the results of [49] about the global maximizers of $D_{\mathcal{E}}$ from the binary i.i.d. families and the binomial families and

how to obtain statements about the local maximizers. Third, Example 4.29 illustrates the difficulties of treating the maximization problems of $D_{\mathcal{E}}$ and $\overline{D}_{\mathcal{E}}$ numerically.

Definition 4.25. Let $\mathcal{X}_1 = \{1, 2\}$. For fixed $n \in \mathbb{N}$ let $\overline{\mathcal{E}_1}$ be the independence model on $\mathcal{X}_1^n$. The symmetric group S_n of $[n]$ operates on $\mathcal{X}_1^n$ via permutations of the factors. Let $\mathcal{X}^{S_n}$ of be the orbit partition of this action, cf. Remark 4.12. The *binary i.i.d. model* $\overline{\mathcal{E}_{\text{iid}}^n}$ of size n is the intersection of $\overline{\mathcal{E}_1}$ with the partition model $\overline{\mathcal{E}_{\mathcal{X}^{S_n}}}$. Equivalently, a probability measure $P \in \overline{\mathcal{E}_1}$ belongs to $\overline{\mathcal{E}_{\text{iid}}^n}$ if and only if it is invariant under the natural action of S_n on $\mathbf{P}(\mathcal{X})$.

The binary i.i.d. model models a collection of independent identically distributed binary random variables. There are $n + 1$ orbits $\mathcal{X}^0, \ldots, \mathcal{X}^n$ of S_n in $\mathcal{X}_1^n$, and $x = (x_i)_{i=1}^n \in \mathcal{X}^k$ if and only if $\sum_{j=1}^n (x_j - 1) = k$. Under the natural bijection from $\overline{\mathcal{E}_{\mathcal{X}^{S_n}}}$ to $\mathbf{P}(\{0, \ldots, n\})$ from Remark 4.13 the i.i.d. model is mapped to the binomial model:

Definition 4.26. The *binomial distribution* with parameters n and p is the probability distribution $P_{n,p}$ on $\mathcal{X} = \{0, \ldots, n\}$ such that

$$P_{n,p}(k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

The *binomial model* of size n is the set of all binomial distributions $\text{Bin}(n) = \{P_{n,p} : 0 \leq p \leq 1\}$. The binomial model $\text{Bin}(2)$ is also called the *Hardy-Weinberg model*.

Both $\overline{\mathcal{E}_{\text{iid}}^n}$ and $\text{Bin}(n)$ are closures of one-dimensional exponential families $\mathcal{E}_{\text{iid}}^n$ and $\text{Bin}(n)^\circ$: The uniform measure is a reference measure for $\mathcal{E}_{\text{iid}}^n$, and a sufficient statistics $a \in \mathbb{R}^{\mathcal{X}_1^n}$ is given by the map $a(x) = \sum_{j=1}(x_j - 1)$ that counts the number of twos among the components of $x \in \mathcal{X}_1^n$. A sufficient statistics $b \in \mathbb{R}^{\mathcal{X}}$ of $\text{Bin}(n)^\circ$ is given by $b = (0, 1, \ldots, n)$, and a reference measure ν is given by $\nu_k = \binom{n}{k}$. By Remark 4.13 maximizing $D(Q \| \text{Bin}(n))$ for $Q \in \mathbf{P}(\{0, \ldots, n\})$ is equivalent to maximizing $D_{\mathcal{E}}(Q)$ subject to $Q \in \overline{\mathcal{E}_{\mathcal{X}^{S_n}}}$. The global maximizers for both exponential families were first computed by Matúš in [49]:

Proposition 4.27. *The maximum of the information divergence from the binary i.i.d. model* $\overline{\mathcal{E}_{\text{iid}}^n}$ *equals* $nh(\frac{1}{n}\lfloor \frac{n}{2} \rfloor, \frac{1}{n}\lceil \frac{n}{2} \rceil)$*, where* $h(p, 1-p)$ *is the entropy of a binary random variable. The set of global maximizers consists of all point measure* δ_x *such that* $x \in \mathcal{X}_1^n$ *satisfies* $\sum_i (x_i - 1) \in \{\lfloor \frac{n}{2} \rfloor, \lceil \frac{n}{2} \rceil\}$.

Proposition 4.28. *The maximum of the information divergence from the binomial model* $\text{Bin}(n)$ *equals* $(n-1)\log(2)$*. If* $n = 2$*, then* δ_1 *and* $\frac{1}{2}(\delta_0 + \delta_2)$ *are the two global maximizers. If* $n > 2$*, then* $\frac{1}{2}(\delta_0 + \delta_n)$ *is the unique global maximizer.*

For the proof of Proposition 4.28 see [49]. The proof of Proposition 4.27 will be given at the end of the section, after the following calculations that give information about the other local maximizers and projection points. For the total number of local maximizers of $\text{Bin}(n)$ and $\mathcal{E}_{\text{iid}}^n$ see Table 4.2.

As discussed in Section 4.1.2, all point measures δ_x such that $0 < a(x) < n$ are local maximizers of $D_{\mathcal{E}^n_{\text{iid}}}$. Similarly, the point measures δ_i for $0 < i < n$ are local maximizers of $D_{\text{Bin}(n)}$. All other local maximizers have support of cardinality two.

Consider first the binomial model. Fix $0 \le i < j \le n$. With the help of the well-known identities

$$\sum_{k=0}^{n} \binom{n}{k} \xi^k = (1+\xi)^n, \qquad \sum_{k=0}^{n} k \binom{n}{k} \xi^k = n\xi(1+\xi)^{n-1},$$
$$\sum_{k=0}^{n} k^2 \binom{n}{k} \xi^k = n\xi(n\xi+1)(1+\xi)^{n-2},$$

the formulas of Section 4.1.2 become

$$\begin{aligned} Z_\xi &= (1+\xi)^n, \\ \lambda(\xi) &= \frac{n\xi(1+\xi)^{n-1} - i(1+\xi)^n}{(j-i)(1+\xi)^n} = \frac{(n-i)\xi - i}{(j-i)(1+\xi)}, \\ 1-\lambda(\xi) &= \frac{j-(n-j)\xi}{(j-i)(1+\xi)} \end{aligned}$$

(note that most formulas in Section 4.1.2 also hold without the normalization $a_n = 1$, and hence they can be applied by setting $a_k = k$ and performing all sums from $k = 0$ to $k = n$; alternatively, the formulas can be applied with $a_k = \frac{k}{n}$).

The inverse of the map $\xi \mapsto \lambda(\xi)$ can be computed explicitly:

$$\xi(\lambda) = \frac{(j-i)\lambda + i}{(n-i) - (j-i)\lambda}.$$

Therefore, $\xi_i = \xi(0) = \frac{i}{n-i}$ and $\xi_j = \xi(1) = \frac{j}{n-j}$. The derivative of $D_{\mathcal{E}}$ equals

$$\frac{\partial}{\partial\lambda} D(P_{\lambda(\xi)} \| P_\xi) = \log \frac{\binom{n}{i}((n-i)\xi - i)}{\binom{n}{j}(j-(n-j)\xi)\xi^{j-i}}.$$

If $j = i+1$, then $\Delta^\circ_{i,j}$ contains no local maximizer. Assume that $j > i+1$. The second derivative of $D(P_\lambda \| P_\xi)$ equals

$$\begin{aligned} \frac{\partial^2}{\partial\lambda^2} D(P_\lambda \| P_\xi) &\sim \sum_{k=0}^{n} (k-i)(k-j) \binom{n}{k} \xi^k \\ &= (1+\xi)^{n-2} \left(n\xi(n\xi+1) - (i+j)n\xi(1+\xi) + ij(1+\xi)^2 \right) \\ &\sim \left(\xi^2(n-i)(n-j) + \xi(n-(i+j)n+2ij) + ij \right) \end{aligned}$$

up to positive factors. If $i = 0$ and $j = n$, then $\frac{\partial^2}{\partial\lambda^2} D(P_\lambda \| P_\xi) \sim -n(n-1) < 0$, and there is one local maximizer of $D_{\mathcal{E}}$ on $\Delta_{0,n}$. From symmetry it follows that this maximizer is $P = \frac{1}{2}(\delta_0 + \delta_n)$, and it projects to the uniform distribution $\frac{1}{2^n}\nu$. Therefore, $D_{\text{Bin}(n)}(P) = (n-1)\log(2)$. This is the global maximum of $D_{\text{Bin}(n)}$ by Proposition 4.28.

If $i = 0$ and $j < n$, then $\frac{\partial^2}{\partial\lambda^2} D(P_\lambda \| P_\xi) \sim n\xi\,(\xi(n-j) - (j-1))$, so $\frac{\partial^2}{\partial\lambda^2} D(P_\lambda \| P_\xi)$ has a unique zero for $0 < \xi < \xi_j$ at $\xi_0 = \frac{j-1}{n-j} < \xi_j$. Therefore,

$$\frac{\partial}{\partial\lambda} D(P_\lambda \| P_\xi) \geq \frac{\partial}{\partial\lambda} D(P_\lambda \| P_{\xi_0})\big|_{\lambda=\lambda(\xi_0)} = \log \frac{n(n-j)^{j-1}}{\binom{n}{j}(j-1)^{j-1}}.$$

The minimum of $\frac{\partial}{\partial\lambda} D_{\mathrm{Bin}(n)}(P_\lambda)$ is negative if and only if

$$f_n(j) := \binom{n}{j} \left(\frac{j-1}{n-j}\right)^{j-1} > n.$$

The function $f_n(j)$ increases monotonically for $j \leq \frac{n}{2}$. For $j > \frac{n}{2}$ the binomial coefficient decreases again, but the fraction $\frac{j-1}{n-j}$ is larger than one, and the binomial coefficient is larger than or equal to n (since $j < n$). Hence there exists a function $j_0(n)$ such that $f_n(j) > n$ if and only if $j \geq j_0(n)$. From the calculation

$$f_n(3) - n = \frac{-n^3 + 12n^2 - 23n}{3(n-3)^2}, \qquad f_n(4) - n = \frac{n^4 + 42n^3 - 285n^2}{8(n-4)^3}$$

one can deduce that

$$j_0(n) = \begin{cases} 3, & \text{if } n < 10, \\ 4, & \text{if } n \geq 10. \end{cases}$$

$D_{\mathrm{Bin}(n)}$ has a local maximum in the relative interior $\Delta^\circ_{0,j}$ of the line segment $\Delta_{0,j}$ if and only if $j \geq j_0(n)$. By symmetry, $D_{\mathrm{Bin}(n)}$ has a local maximum in $\Delta^\circ_{i,n}$ if and only if $i \leq n - j_0(n)$.

Finally, consider the case $0 < i < i+1 < j < n$. In this case, the second derivative $\frac{\partial^2}{\partial\lambda^2} D_{\mathrm{Bin}(n)}(P_\lambda)$ vanishes if and only if $\xi \in \{\xi_+, \xi_-\}$, where

$$\xi_\pm = \frac{(i+j-1)n - 2ij \pm \sqrt{((j-i)^2 - 2(i+j) + 1)n^2 + 4ijn}}{2(n-i)(n-j)}. \tag{4.15}$$

The term $w(i,j,n) := ((j-i)^2 - 2(i+j) + 1)n^2 + 4ijn$ under the root may assume positive and negative values. Writing $j = i + \delta$ it can be rewritten as

$$w(i, i+\delta, n) = n(n\delta^2 + (4i - 2n)\delta + 4i^2 - 4in + n).$$

For fixed i and n the function $w(i, i+\delta, n)$ is monotonically increasing in δ. Let $\delta_0(i,n)$ be the smallest value of $\delta \in \mathbb{N}$ such that $w(i, i+\delta, n) \geq 0$. Then $w(i,j,n) \geq 0$ if and only if $(j-i) \geq \delta_0(i,n)$.

Assume that $w(i,j,n) \geq 0$. Since $i + j - 1 \geq 2i$ and $n > j$ it follows that $(i+j-1)n - 2ij > 0$. Therefore, both solutions ξ_+ and ξ_- are positive. From $4ijn < 4jn^2$ it follows

$$\xi_+ < \frac{(i+j-1)n - 2ij + (j-i+1)n}{2(n-i)(n-j)} = \xi_j.$$

Similarly, $4ijn < 4in^2$ implies $\xi_- > \xi_i$. Therefore, $\frac{\partial}{\partial\lambda} D_{\mathrm{Bin}(n)}(P_\lambda)$ shows the following behaviour: For $\xi \to \xi_i$ it diverges to $-\infty$, as discussed in Section 4.1.2. Then it increases until it reaches a local maximum at ξ_-. From ξ_- it decreases down to its local minimum at ξ_+. From ξ_+ to ξ_j it increases again monotonically, diverging to $+\infty$ at ξ_j. Hence $D_{\mathrm{Bin}(n)}(P_\lambda)$ has zero or one local maximizer in $\Delta^\circ_{i,j}$, and the second case occurs if and only if $\frac{\partial}{\partial\lambda} D_{\mathrm{Bin}(n)}(P_{\lambda(\xi_-)}) > 0 > \frac{\partial}{\partial\lambda} D_{\mathrm{Bin}(n)}(P_{\lambda(\xi_+)})$.

Now consider the binary i.i.d. model. Let $x, y \in \mathcal{X}_1^n$. For any $0 < \lambda < 1$ let $\tilde{P}_\lambda = (1-\lambda)\delta_x + \lambda\delta_y$. By the results of Section 4.2 the rI-projection of $\tilde{P}_\lambda$ onto the partition model $\mathcal{E}_{\mathcal{X}^{S_n}}$ equals

$$(\tilde{P}_\lambda)_{\mathcal{X}^{S_n}} := (1-\lambda)\frac{1}{\binom{n}{a(x)}} \sum_{z:a(z)=a(x)} \delta_z + \lambda \frac{1}{\binom{n}{a(y)}} \sum_{z:a(z)=a(y)} \delta_z.$$

It is mapped by the natural bijection $\phi_* : \mathcal{E}_{\mathcal{X}^{S_n}} \to \mathbf{P}(\mathcal{X}^{S_n})$ to $P_\lambda = (1-\lambda)\delta_{a(x)} + \lambda\delta_{b(x)}$. In particular, if $\xi = \xi(\lambda)$ is defined as above, then the rI-projection of $\tilde{P}_\lambda$ onto $\mathcal{E}^n_{\mathrm{iid}}$ equals $\tilde{P}_\xi$, where

$$\tilde{P}_\xi(x) = \frac{1}{Z_\xi}\xi^{a(x)} \quad \text{for all } x \in \mathcal{X}_1^n.$$

Hence

$$\begin{aligned} D(\tilde{P}_\lambda \| \mathcal{E}^n_{\mathrm{iid}}) &= D(\tilde{P}_\lambda \| (\tilde{P}_\lambda)_{\mathcal{X}^{S_n}}) + D((\tilde{P}_\lambda)_{\mathcal{X}^{S_n}} \| \mathcal{E}^n_{\mathrm{iid}}) \\ &= (1-\lambda)\log\binom{n}{a(x)} + \lambda\log\binom{n}{a(y)} + D(P_\lambda \| \operatorname{Bin}(n)). \end{aligned}$$

Therefore,

$$\begin{aligned} \frac{\partial}{\partial\lambda} D(\tilde{P}_\lambda \| \mathcal{E}^n_{\mathrm{iid}}) &= \frac{\partial}{\partial\lambda} D(P_\lambda \| \operatorname{Bin}(n)) + \log\frac{a(x)!(n-a(x))!}{a(y)!(n-a(y))!}, \\ \frac{\partial^2}{\partial\lambda^2} D(\tilde{P}_\lambda \| \mathcal{E}^n_{\mathrm{iid}}) &= \frac{\partial^2}{\partial\lambda^2} D(P_\lambda \| \operatorname{Bin}(n)), \end{aligned}$$

and the following statements can be made from the above discussion of $D(P_\lambda \| \operatorname{Bin}(n))$:

- If $a(y) = a(x)$, then there is no local maximizer in $\Delta^\circ_{x,y}$ by Lemma 3.5.
- If $a(y) = a(x) + 1$, then there is no local maximizer in $\Delta^\circ_{x,y}$ by Theorem 4.2.
- If $a(x) = 0$ and $a(y) = n$, then $P = \frac{1}{2}(\delta_x + \delta_y)$ is the unique local maximizer on $\Delta_{x,y}$. It rI-projects to the uniform distribution and satisfies $D(P\|\mathcal{E}^n_{\mathrm{iid}}) = (n-1)\log(2)$.
- If $a(x) = 0$ and $0 < a(y) < n$, then

$$\frac{\partial}{\partial\lambda} D_{\mathcal{E}^n_{\mathrm{iid}}}(\tilde{P}_\lambda) \geq \log n \left(\frac{n-a(y)}{a(y)-1}\right)^{a(y)-1},$$

n:	4	5	6	6	7	7	7	7	8	8	8	8	8	8	9	9	9	9	9	9	9	9	9
i:	1	1	1	1	1	2	1	1	1	2	1	2	1	1	1	2	3	1	2	1	2	1	1
j:	3	4	4	5	4	5	5	6	4	5	5	6	6	7	4	5	6	5	6	6	7	7	8
Bin:	1^*	3	3	3	3	3	3	3	1	2	3	3	3	3	1	1	1^*	3	3	3	3	3	3
$\mathcal{E}^n_{\text{iid}}$:	1^*	3	1	3	1	3	3	3	1	1	1	3	3	3	1	1	1^*	1	1	3	3	3	3

Table 4.1.: The number of projection points with specific supports in $\text{Bin}(n)$ and $\mathcal{E}^n_{\text{iid}}$.

n:	2	3	4	5	6	7	8	9
$\text{Bin}(n)$:	2	3	6	10	15	21	24	31
$\mathcal{E}^n_{\text{iid}}$:	3	7	15	66	111	967	1623	3235

Table 4.2.: The total number of local maximizers of $D_{\mathcal{E}}$ for $\text{Bin}(n)$ and $\mathcal{E}^n_{\text{iid}}$.

and the minimum is negative if and only if

$$g_n(a(y)) := \left(\frac{a(y)-1}{n-a(y)}\right)^{a(y)-1} > n.$$

The function g_n is monotonically increasing. Hence there exists a function $j_1(n)$ such that $D_{\mathcal{E}^n_{\text{iid}}}$ has a local maximum in $\Delta^\circ_{x,y}$ if and only if $a(y) \geq j_1(n)$. From $g_n(\lceil \frac{n}{2} \rceil) \leq 1 \leq n$ it follows that $j_1(n) > \lceil \frac{n}{2} \rceil$. Let $1 > \beta > \frac{1}{2}$. Then $g_n(\beta(n-1)+1) - n = \left(\frac{1}{\beta} - 1\right)^{-\beta(n-1)} - n \to \infty$ as $n \to \infty$, and so

$$\lim_{n\to\infty} \frac{j_1(n)}{n} = \lim_{n\to\infty} \frac{j_1(n)-1}{n-1} = \frac{1}{2}.$$

The following table gives the value of $j_1(n)$ for $4 \leq n \leq 13$:

n:	4	5	6	7	8	9	10	11	12	13
$j_1(n)$:	4	4	5	5	6	6	7	7	8	9

- Similarly, if $0 < a(x) < n = a(y)$, then there is a local maximum in $\Delta^\circ_{x,y}$ if and only if $a(x) \leq n - j_1(n)$.
- If $0 < a(x) < a(y) < n$, then $D_{\mathcal{E}^n_{\text{iid}}}(\tilde{P}_\lambda)$ has zero or one local maximizer in $\Delta^\circ_{x,y}$. The second case occurs if and only if $w(a(x), a(y), n) > 0$ and $\frac{\partial}{\partial\lambda} D_{\mathcal{E}^n_{\text{iid}}}(P_{\lambda(\xi_-)}) > 0 > \frac{\partial}{\partial\lambda} D_{\mathcal{E}^n_{\text{iid}}}(P_{\lambda(\xi_+)})$, where $\xi_\pm$ are defined in (4.15).

Table 4.1 shows the number of projection points with support of cardinality two for $\text{Bin}(n)$ and $\mathcal{E}^n_{\text{iid}}$ for $n \leq 9$. The fourth row gives the number of projection points in $\Delta^\circ_{i,j}$ for $\text{Bin}(n)$, and the fifth row gives the number of projection points in $\Delta^\circ_{x,y}$ for $\mathcal{E}^n_{\text{iid}}$ whenever $i = a(x)$ and $j = a(y)$. Only the values for $i \leq n - j$ are printed; the others can be reconstructed by using the symmetry $(i,j) \leftrightarrow (n-j, n-i)$. If a triple (n,i,j) with $i \leq n - j$ is missing, then $w(i,j,n) < 0$, and $D_{\text{Bin}(n)}$ and $D_{\mathcal{E}^n_{\text{iid}}}$ are both convex on the corresponding segment. Hence there is one projection point, which is

not a local maximum. An asterisk marks the case that $w_{i,j,n} = 0$. In this case there is exactly one projection point P, and at this projection point the second derivative of $D_{\mathcal{E}}$ along $\Delta_{i,j}$ resp. $\Delta_{x,y}$ vanishes at P, where $\mathcal{E} = \mathrm{Bin}(n)$ or $\mathcal{E} = \mathcal{E}^n_{\mathrm{iid}}$. This projection point must be a local minimum of the information divergence restricted to $\Delta_{i,j}$, and so the third derivative also vanishes. If there are two projection points, then only one is a local minimum of $D_{\mathcal{E}}|_{\Delta_{i,j}}$, and the other is a saddle point, where the first and second derivatives vanish. If there are three projection points, then there are two local minima and one local maximum of $D_{\mathcal{E}}|_{\Delta^\circ_{i,j}}$.

From these results, the total number of local maximizers of $\mathrm{Bin}(n)$ and $\mathcal{E}^n_{\mathrm{iid}}$ can be counted. The result is shown in Table 4.2.

Proof of Proposition 4.27. If n is even, then the statement follows directly from Proposition 4.4. Otherwise, assume that $P \in \Delta_{x,y}$ is a global maximizer of $D_{\mathcal{E}^n_{\mathrm{iid}}}$. By symmetry and Proposition 4.4 one may assume that $a(x) = 0$ and $a(y) = \lceil \frac{n}{2} \rceil$. By the calculations above $a(y) < j_1(n)$, and hence there is no critical point of $D_{\mathcal{E}^n_{\mathrm{iid}}}$ in $\Delta^\circ_{x,y}$. Therefore, P is an endpoint of $\Delta_{x,y}$, and so $P = \delta_y$, since $\delta_x \in \overline{\mathcal{E}^n_{\mathrm{iid}}}$. □

This section ends with a detailed discussion of one of the critical points of $\mathrm{Bin}(4)$, which shows how awkward the maximization problem can be.

Example 4.29. By Table 4.1 there is a projection point P of $\mathrm{Bin}(4)$ in $\Delta_{1,3}$ at which the first three derivatives of $D_{\mathrm{Bin}(4)}$ vanish. By symmetry $P = \frac{1}{2}(\delta_1 + \delta_3)$, and P rI-projects to $\frac{1}{8}\nu = \frac{1}{8}(1,3,3,1)$. Let $u = \Psi_{\mathcal{E}}(P) = \frac{1}{6}(2,-3,2,-3,2) \in \partial \mathbf{U}_{\mathcal{N}}$. The map

$$(x,y) \mapsto u + \frac{1}{8}(x+y, -2x, -2y, 2x, y-x)$$

is a linear parametrization of a neighbourhood of u in $\partial \mathbf{U}_{\mathcal{N}}$. The Hessian of $(x,y) \mapsto \overline{D}_{\mathcal{E}}(x,y)$ at $(0,0)$ equals

$$\begin{pmatrix} 0 & 0 \\ 0 & -\frac{1}{3} \end{pmatrix}.$$

Because of $\frac{\partial^3}{\partial x^3} \overline{D}_{\mathcal{E}}(u(x,y))\big|_{x=y=0} = 0$ and $\frac{\partial^4}{\partial x^4} \overline{D}_{\mathcal{E}}(u(x,y))\big|_{x=y=0} = -\frac{3}{8}$ the restriction of $\overline{D}_{\mathcal{E}}$ to any line through u has a local maximum at u. But u is not a local maximum, because of Theorem 3.16. This can be seen directly as follows: Consider the function $f(x) = \overline{D}_{\mathcal{E}}(u(x, -\frac{3}{8}x^2))$. One computes $f'(0) = f''(0) = f'''(0) = 0$ and $f''''(0) = \frac{3}{2}$. Therefore, the function $\overline{D}_{\mathcal{E}}(u(x,y))$ restricted to the curve $y + \frac{3}{8}x^2 = 0$ has a local minimum at the origin, and so u is only a saddle point of $\overline{D}_{\mathcal{E}}$.

The existence of real functions that have a local maximum on every line that passes through some fixed point x but that do not have a local maximum at x is usually the subject of an exercise in an undergraduate analysis course. Nevertheless, it is surprising that this phenomenon appears here not in some constructed example, but in a rather natural case.

5. Applications and Outlook

This chapter is dedicated to applications of the mathematical theory developed in the previous chapters. The three sections of this chapter discuss three different applications. The first two sections are related to the maximization of the information divergence; the third section is related to the projection points. Sections 5.1 and 5.2 also state some open problems and a conjecture that indicate future research directions.

Section 5.1 reviews the original motivation to study the maximizers of $D_{\mathcal{E}}$. This section does not contain new mathematical results. It explains how the results of this thesis apply and which extensions are necessary to make these results useful for the study of learning in neural networks. In particular, the most important next step is to study the maximization of the information divergence under suitable constraints. This constrained optimization problem is well-known in the theory of channel capacities.

Section 5.2 discusses an application to machine learning. An important topic in machine learning is the study of algorithms for learning a probability distribution by fitting experimental data to a statistical model (in many cases an exponential family, or a subset of an exponential family). The idea pursued in this section is to search for exponential families that can approximate arbitrary empirical probability distributions well, in the sense of a low maximum value of $D_{\mathcal{E}}$. The section defines two appropriate notions of optimality and identifies the homogeneous partition models from Section 4.2 as a class of models that contains many optimal models.

Section 5.3 studies the asymptotic behaviour of the empirical information divergence from an exponential family. This behaviour changes qualitatively if the underlying probability distribution that generates the empirical data is a projection point. The goal of this section is to study and characterize this behaviour.

5.1. Principles of learning, complexity measures and constraints

Artificial neural networks are mathematical models that have been developed to study biological neural networks in brains and nervous systems of living beings. Neural networks have also proved to be useful for solving decision tasks and other computational problems. See [37] for an introduction. Just like the biological prototypes, neural networks consist of discrete units, called *nodes* or *neurons*. Each neuron receives an input signal from other neurons to which it is connected with a certain connection strength. From this input, an output signal is computed and passed on to other neurons. There are also sensor neurons that receive input from the outside. The result of

a computation is read out from the state of output neurons. Such a network is called a *feed-forward neural network* if the neurons can be ordered such that no neuron receives input from subsequent neurons.

Depending on the application in mind the different components of the neural network can be modelled with more or less detail. In the simplest case the neurons are binary units that compute a Boolean function of their inputs (e.g. a threshold function of the sum of the inputs, weighted by the connection weights). This abstraction goes back to McMurdoch and Pitts [54]. In more detailed models the units themselves are described by one or more real variables. Moreover, the models may have discrete or continuous time evolution. For single neurons there are differential equations that can quantitatively reproduce the behaviour, e.g. the Hodgkin-Huxley model [39] and related models. When simulating large-scale networks the single neurons have to be modelled in a much simpler way.

An important feature of neural networks is that the connection strength of the nodes is usually not hard-coded in the beginning. Instead, the network is trained for a specific task with the help of training data, and it may even adapt to a continuously changing environment. It is not necessary for the programmer to implement a complicated algorithm into the network: If the structure of the network is sufficiently general and adapted to the problem to be solved, then the network is able to learn a solution strategy by itself. Therefore, the task of the programmer is to choose the design of the network appropriately for the task at hand and to specify a suitable learning algorithm.

One of the oldest and easiest learning rules for neural networks is Hebb's rule that was postulated by the psychologist Hebb on theoretical grounds. It states that if two neurons often fire together, then the connection strength between these neurons increases. Later, this rule could be confirmed experimentally for some neural systems (and disproved for others). For example, the phenomenon of *spike timing dependent plasticity* (see [12] and references therein) is a variant of Hebb's rule: Plasticity means that the connection strength has the possibility to adapt, and this adaptation depends on the relative timing of excitations of the neurons. If neuron A often fires shortly before neuron B, then the connection from A to B is strengthened and the connection from B to A is weakened, and vice versa.

Motivated by Hebb's rule Linsker proposed an information theoretic learning principle in [47], called the *infomax principle*. Linsker was interested in unsupervised learning of perceptive systems. Such systems usually have a feed-forward structure and consist of distinct layers. The first layer receives input signals from sensors, and it is assumed that the distribution of the signals is according to some fixed stationary distribution. Subsequent layers receive input from their precedent layer. Each layer uses its input to compute some output that is then passed to the next layer. The infomax principle states:

- The system learns in such a way that the mutual information between the input and the output of each layer is maximized.

In [6] Ay proposed a related principle, the *IMI principle*. This principle does not

need the structuring of the network into layers; however, it is still assumed that some of the neurons receive external input. The IMI principle states:

- The system learns in such a way that the multiinformation between the neurons is maximized.

Ay shows that for a two-layer feed-forward network a gradient ascent with respect to the multiinformation is equivalent to a Hebb-like rule. See [6] for a comparison of the infomax and IMI principles.

The two information theoretic principles stated so far only involve the stationary distribution of the neurons. In [9] Ay and Wennekers generalized these ideas and formulated a learning principle that involves the dynamics of the process in the form of a Markov transition kernel: Consider a neural network with state space $\mathcal{X}_0$, and let $\mathcal{X} = \mathcal{X}_0 \times \mathcal{X}_0$. Write $X, Y : \mathcal{X} \to \mathcal{X}_0$ for the two canonical projection maps. The two factors model the present and the future of the system. Let $P \in \mathbf{P}(\mathcal{X})$. Assume that the marginal distributions of X and Y agree. In this case, the system is in a stationary state. The dynamics is encoded in the conditional probability distributions $P(Y = y|X = x)$. In this setting, Ay and Wennekers propose to use the conditional information divergence

$$\sum_{x \in \mathcal{X}_0} P(x) \sum_{y \in \mathcal{X}_0} P(y|x) \log \frac{P(y|x)}{Q(y|x)} \tag{5.1}$$

to compare different stationary distributions and different dynamics, leading to the concept of *temporal infomax* [72].

A second motivation to investigate the maximizers of the information divergence is the study of complexity measures. A widespread paradigm claims that a system is complex if it cannot be described by looking only at smaller subsystems. Ay proposed to formalize this idea in [4] by measuring the distance of the state of the system from some suitably defined product state. If the system is described by Markov kernels, then (5.1) can be used to quantify the distance of some kernel from the set of all factorizable kernels, see [4] for the details. If the state of the system is described by probability distributions, then choosing the information divergence as a distance measure and the independence model $\overline{\mathcal{E}_1}$ as the set of product states yields the multiinformation. A related idea was pursued by Ay, Olbrich, Bertschinger and Jost in [8]: For a composite system $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ consider the interaction models $\mathcal{E}_1 \subset \mathcal{E}_2 \subset \cdots \subset \mathcal{E}_{n-1}$ as defined in Section 2.4. For any vector $\alpha \in \mathbb{R}^{n-1}$ let

$$C_\alpha(P) := \sum_{k=1}^{n-1} \alpha(k) D(P \| \mathcal{E}_k).$$

The different terms $D(P\|\mathcal{E}_k)$ measure how well the distribution P can be described by interactions of only k of the subsystems. The results of this thesis are directly applicable to the study of C_α whenever α has only one non-vanishing component. Another approach is taken in [42]. Based on the idea that a system is complex if

any faithful description of the behaviour involves all levels, all components $D(P\|\mathcal{E}_k)$ are considered together. By the Pythagorean identity $D(P\|\mathcal{E}_k) = \sum_l^{n-1} D(P_{l+1}\|\mathcal{E}_l)$, where P_l is the rI-projection of P onto $\mathcal{E}_l$ (this decomposition was first studied by Amari [2]). This motivates to consider the *interaction complexity vector* $IC(P) := (D(P_{l+1}\|\mathcal{E}_l))_{l=1}^{n-1}$. Its lth component can be interpreted as the proportion of the system that can be understood with the help of subsystems of size $l+1$, but not through the subsystems of size l. These ideas are related to the *TSE-complexity*, introduced by Tononi, Sporns and Edelman in [70], see [8] for a comparison.

The mathematical results of this thesis are not directly applicable to the study of the infomax and IMI principles in a biological context, because in these principles the maximization usually takes place under certain internal and external constraints: External constraints are given by the distribution of the input signals that the system receives. Internal constraints are given by the network structure itself. For example, the interaction between the neurons consists of pair interactions. Furthermore, depending on the model, a single neuron cannot compute arbitrary functions from its input. Other internal constraints may stipulate which connections are inhibiting and which connections are exciting.

Two kinds of constraints appear naturally in the above considerations: Linear constraints can be used to model external constraints to a network in the form of marginal distributions. Linear constraints also appear in the study of symmetries, cf. Remark 4.12. Exponential constraints of the form $Q \in \mathcal{E}'$ for some exponential family $\mathcal{E}'$ can be used to model certain internal constraints. For example, $\mathcal{E}'$ could be taken to be a submodel of the pair interaction model. Such exponential constraints also appear in the study of the components IC_l of the complexity vector $IC(P)$: The maximization of a single component $IC_l(P) = D(P_{l+1}\|\mathcal{E}_l)$ corresponds to the maximization of $D(Q\|\mathcal{E}_l)$ subject to $Q \in \mathcal{E}_{l+1}$. These two kinds of constraints were introduced in the framework of the IMI principle in [6]. In the dynamical framework, constraints were studied in [72]. A special case of the constrained optimization of $D_{\mathcal{E}}$ appeared in this thesis in Remark 4.13 and in Section 4.4: If $\mathcal{E}'$ is a partition model and if $\mathcal{E} \subseteq \mathcal{E}$, then the constrained optimization of $D_{\mathcal{E}}(Q)$ subject to $Q \in \overline{\mathcal{E}'}$ is equivalent to an unconstrained optimization problem. More generally, it is easy to see that the same holds true if $\mathcal{E}'$ is any convex exponential family containing $\mathcal{E}$ as a subfamily (but not necessarily containing the uniform distribution).

Even though the results in this thesis are not directly applicable to the learning principles, one can argue that they are still relevant to the general problem: First, the unconstrained problem gives bounds on the constrained problem. Second, if the learning principles stated above are sufficiently important for the fitness of the biological system, then the biological system should evolve such that the constraints on the system do not obstruct the maximization of the corresponding entropic quantity. For example, the sensors should evolve such that the external constraints do not restrict the learning of the system, and the internal structure should evolve such that the maximizers of the unconstrained optimization principle are still reachable. This line of arguments was first proposed by Ay in [6]. It also relates learning principles and complexity measures: It is a well-known phenomenon that the evolution of biologi-

cal systems creates complex structures, see for example [62] and [66] and references therein. In the light of the learning principles and the complexity measures discussed above a possible interpretation is the following: The evolution of neural networks creates complex structures, in the sense that the resulting neural networks can exhibit complex behaviour. The fine-tuning of the connection strengths is achieved by learning and depends on the external constraints given by the environment. Both phases, the evolution phase and the learning phase, are imitated in the theory of artificial neural networks: First, the network structure is chosen (by principle considerations or by evolutionary algorithms), and then the network learns for a given task.

If the motivation is not only to understand biological neural networks but to construct artificial neural networks in order to solve certain tasks, then the learning principles suggest to construct such systems in a way that they can reach the maximizers. This leads to the following mathematical problem:

- Given an exponential family $\mathcal{E}$, find a low-dimensional statistical model $\mathcal{E}'$ that contains all (local or global) maximizers of $D_{\mathcal{E}}$ in it closure.

One may then try to construct a neural network that can generate all probability distributions in $\mathcal{E}'$. If the IMI principle is appropriate to the problem to be solved, then such neural networks should perform well. In [52] Matúš and Ay show that for any exponential family $\mathcal{E}$ of dimension d there exists an exponential family $\mathcal{E}'$ of dimension $3d+2$ that contains all local maximizers of $D_{\mathcal{E}}$. Their idea is to construct $\mathcal{E}'$ such that for any subset $\mathcal{Z} \subseteq \mathcal{X}$ of cardinality $|\mathcal{Z}| \leq \dim \mathcal{E} + 1$ and any $Q \in \mathcal{E}$ the truncation $Q^{\mathcal{Z}}$ is contained in $\overline{\mathcal{E}'}$, cf. Lemma 3.5. In [7] Ay and Knauf show that the global maximizers from those independence models $\mathcal{E}_1$ that satisfy (1.1) lie in the closure of the exponential family of pure pair-interactions $\mathcal{E}_{(2)}$. If $d = \dim \mathcal{E}_1$, then the dimension of $\mathcal{E}_{(2)}$ grows approximately quadratic in d, and hence it grows faster than the bound $3d+2$ of [52]. On the other hand, $\mathcal{E}_{(2)}$ has the advantage that it has a nice interpretation. Another strategy could be to search for a statistical model $\mathcal{M}$ containing the set $\mathrm{K}_{\mathcal{E}}$ of kernel distributions in its closure, see Section 3.2.

Of course, for a real biological system there will be other factors besides the learning principles stated above that influence its evolution. In general, it is too bold to assume that the learning system may reach the global maximizers of the unconstrained optimization problem despite the internal and external constraints. Therefore, the generalization of the results of this thesis to the constrained maximization of $D_{\mathcal{E}}$ is an important problem for the future.

Under suitable constraints the maximization of $D_{\mathcal{E}}$ also makes sense in the case where $\mathcal{X}$ is infinite. This makes it possible to study neurons with a continuous output. The constrained maximization of $D_{\mathcal{E}}$ also appears in information theory in the study of (noisy) channels: Consider a (finite or infinite) system $\mathcal{X} = \mathcal{X}_{\mathrm{in}} \times \mathcal{X}_{\mathrm{out}}$ of two random variables $X : \mathcal{X} \to \mathcal{X}_{\mathrm{in}}$ and $Y : \mathcal{X} \to \mathcal{X}_{\mathrm{out}}$, called the *input* and the *output* of the channel. A channel can be seen as a map c from a set $\mathbf{P}_{\mathrm{in}} \subseteq \mathbf{P}(\mathcal{X}_{\mathrm{in}})$ of possible input distributions to the joint distributions $\mathbf{P}(\mathcal{X})$. The input distribution corresponds to a stationary stochastic source of information, and the joint distribution determines the relation between the output and the input. In the easiest case, it suffices to give

a conditional distribution $P(\cdot|x) \in \mathbf{P}(\mathcal{X}_{\text{out}})$ for any $x \in \mathcal{X}_{\text{in}}$. For example, a noisy Gaussian channel can be specified by $\mathcal{X}_{\text{in}} = \mathcal{X}_{\text{out}} = \mathbb{R}$ and $Y = X + \sigma\epsilon$, where $\sigma \in \mathbb{R}_{\geq}$ and ϵ is normally distributed with variance 1. In this case the conditional distributions $P(\cdot|x)$ are also normally distributed with mean x and variance σ^2. The *capacity* of such a channel is defined as

$$C = \sup_{P \in \mathbf{P}_{\text{in}}} I_{c(P)}(X;Y),$$

where $I_{c(P)}(X;Y)$ is the mutual information of input and output, computed with respect to the probability distribution $c(P)$. Depending on the subset $\mathbf{P}_{\text{in}}$, the value of C may be finite or infinite. As an example, consider the noisy Gaussian channel. If $\mathbf{P}_{\text{in}}$ equals the set of all probability distributions with bounded variance, then the capacity is achieved by a Gaussian input distribution, as Shannon showed in [63]. If $\mathbf{P}_{\text{in}}$ equals the set of probability distributions on a compact interval $[-a, a] \subset \mathbb{R}$, then the capacity is achieved by an input distribution with finite support, see [64]. See [17] and references therein for more results on the capacity of channels.

This gives the following interpretation of the infomax principle: Each layer in a feed-forward neural network can be seen as a channel, and learning optimizes these channels. In this context, the mathematical problem of finding statistical models that can approximate the local or global maximizers of the mutual information corresponds to a search for parametrized families of channels that can approximate optimal channels.

5.2. Optimally approximating exponential families

In this section the following question will be discussed:

- Fix a real number $D > 0$ and a partial order on the exponential families. Which exponential families are minimal among all exponential families $\mathcal{E}$ satisfying $\max D_{\mathcal{E}} \leq D$? What is the answer to this question under further constraints on $\mathcal{E}$?

There are at least two partial orders of interest:

(i) The partial order induced by the dimensions of the exponential families.

(ii) The partial order by inclusion.

The partial order (i) is particularly important for applications, since the dimension of an exponential family is one of the most important invariants that determine the complexity of all computations. The partial order (ii) can be seen as a "local relaxation": A candidate exponential family $\mathcal{E}$ is only compared to "similar" exponential families, contained in $\mathcal{E}$.

Definition 5.1. Let $\mathcal{X}$ be a finite set and let $\mathcal{H}$ be a set of exponential families. An exponential family $\mathcal{E} \in \mathcal{H}$ is called *inclusion D-optimal among $\mathcal{H}$* for some $D \geq \max D_{\mathcal{E}}$ if every exponential family $\mathcal{E}' \in \mathcal{H}$ strictly contained in $\mathcal{E}$ satisfies $\max D_{\mathcal{E}} \leq D < \max D(\cdot\|\mathcal{E}')$. An exponential family $\mathcal{E} \in \mathcal{H}$ is called *dimension D-optimal among $\mathcal{H}$* if every exponential family $\mathcal{E}' \in \mathcal{H}$ of smaller dimension satisfies $\max D_{\mathcal{E}} \leq D < \max D(\cdot\|\mathcal{E}')$. Exponential families that are inclusion or dimension D-optimal among $\mathcal{H}$ for some D are also called *inclusion* or *dimension optimal among $\mathcal{H}$*, without reference to D. If $\mathcal{H}$ equals the set of all exponential families, then the reference to $\mathcal{H}$ may be omitted in all definitions. Let

$$D_{N,k}(\mathcal{H}) = \min\{\max D_{\mathcal{E}} : \mathcal{E} \in \mathcal{H} \text{ is an exponential family of dimension } k \text{ on } [N]\}.$$

As an example, the set $\mathcal{H}$ may be the set of hierarchical models, or the set $\mathcal{H}_{\mathbf{1}}$ of exponential families containing the uniform distribution. Obviously, any dimension optimal model is also inclusion optimal. The converse statement does not hold, see Example 5.4 below.

A D-optimal exponential family $\mathcal{E}$ can approximate arbitrary probability measures well, up to a maximal divergence of D. Yaroslav Bulatov proposed to use such exponential families in machine learning (personal communication). A well-known learning principle, sometimes called *minimax principle*, suggests the following algorithm for learning a probability distribution from a set of samples $x_1, \ldots, x_m \in \mathcal{X}$: Let $\mathcal{F}$ be a finite collection of subsets of $\mathbb{R}^{\mathcal{X}}$. Elements of $\mathbb{R}^{\mathcal{X}}$ are also called *features* in this context, so any $F \in \mathcal{F}$ is a candidate feature set. For any $F \subseteq \mathbb{R}^{\mathcal{X}}$ denote by $\mathbb{R}F$ the vector space generated by F, and let $\mathcal{E}_F$ be the exponential family with tangent space $\mathbb{R}F/\mathbb{R}\mathbf{1}$ and uniform reference measure. Denote the empirical distribution over the set of samples by $\hat{P} = \frac{1}{m}\sum_{k=1}^{m} \delta_{x_k}$. The algorithm can be sketched as follows:

1. Start with a subset F^0 of $\mathcal{F}$.
2. In the kth iteration select a set of candidate feature sets $\tilde{\mathcal{F}}^k \subseteq \mathcal{F}$ (this may depend on F^{k-1}).
3. For each $F \in \tilde{\mathcal{F}}^k$ find an estimate D_F^k for $D(\hat{P}\|\mathcal{E}_{F^{k-1}\cup F})$.
4. Let $F^{k+1} = F^k \cup \operatorname{argmin}\left\{D_F^k : F \in \tilde{\mathcal{F}}^k\right\}$.
5. Iterate until the fit is good enough, e.g. until $\min_{F\in\tilde{\mathcal{F}}^k} D_F^k$ is small enough.

There are different possibilities how to fill in the details. For example, the question when to stop in order to prevent overfitting is a version of the difficult statistical problem of model selection. The name of the algorithm refers to the fact that the elements of the exponential family maximize the entropy, subject to constraints given by the expectation values of the features, cf. Theorem 2.16 (*iii*). This is motivated by Jaynes' principle, which states that the maximum entropy estimate is the most objective way to incorporate knowledge about the true underlying probability distribution [40]. On the other hand, the exponential family is chosen such that it minimizes the information

divergence. The name was proposed in the 1997 paper [74] by Zhu, Wu and Mumford, who discuss applications to texture modelling. In the same year, Della Pietra, Della Pietra and Lafferty presented a similar "feature induction algorithm," which they apply to the problem of automatic word classification in natural languages [24]. Both papers assume that the set of possible features is given a priori; they do not specify a way to deduce the candidate features without expert knowledge. The tangent spaces of optimal exponential families may be natural candidates for features, if no or little expert knowledge is available. Some suggestions for the choice of possible candidate features are given at the end of this section.

One motivation to restrict the class $\mathcal{H}$ of exponential families is that the learning system may not be able to represent arbitrary exponential families. Another motivation is given by Jaynes' principle, which suggests to use the class $\mathcal{H}_\mathbf{1}$ of exponential families with uniform reference measure.

Remark 5.2. Remark 3.15 says that $\max D_\mathcal{E} \geq \log(2)$ for all exponential families $\mathcal{E} \neq \mathbf{P}(\mathcal{X})^\circ$. Therefore D-optimality is only interesting for $D \geq \log(2)$. The case $D = \log(2)$ was already studied in Section 4.3. The result is that $D_{N,k} = \log(2)$ for $\lceil \frac{N}{2} \rceil - 1 \leq k < N$. This condition is equivalent to $\lceil \frac{N}{k+1} \rceil = 2$. Many $\log(2)$-dimension optimal exponential families are partition exponential families.

Example 5.3. All zero-dimensional exponential families are dimension-optimal. By Section 4.1.1, if $\mathcal{E} = \{\nu\}$, then

$$\max D_\mathcal{E} = \max\{-\log(\nu_x) : x \in \mathcal{X}\} \geq \log|\mathcal{X}|.$$

Therefore, $D_{N,1} = \log(N)$, and $\mathcal{E}$ is D-optimal if and only if $\nu_x \geq e^{-D}$ for all $x \in \mathcal{X}$. Zero-dimensional exponential families are the dimension D-optimal exponential families for $D \geq \log|\mathcal{X}|$. In general, they are not the only inclusion D-optimal exponential families, see Example 5.4.

Example 5.4. Let $\mathcal{X} = \{1, 2, 3\}$. Any zero-dimensional exponential family $\mathcal{E} = \{\nu\}$ satisfies $\max D_\mathcal{E} \geq \log(3)$. Therefore, if $\log(2) \leq D < \log(3)$, then the dimension D-optimal exponential families are one-dimensional. The case $D = \log(2)$ was already discussed in Proposition 4.20. The general case can be treated similarly:

The normal space $\mathcal{N}$ of any one-dimensional exponential family $\mathcal{E}$ is spanned by a single element u, which can be taken to be normalized, such that $\partial\mathbf{U}_\mathcal{N} = \{\pm u\}$. By Example 3.13 the set of local maximizers of $D_\mathcal{E}$ equals $\{u^+, u^-\}$. Let $P_\mathcal{E} = (u^+)_\mathcal{E} = (u^-)_\mathcal{E}$, then $P_\mathcal{E} = \mu u^+ + (1-\mu)u^-$ for some $0 < \mu < 1$. Hence $D_\mathcal{E}(u^+) = -\log\mu$ and $D_\mathcal{E}(u^-) = -\log(1-\mu)$. It follows that $\mathcal{E}$ is dimension D-optimal if and only $\exp(-D) \leq \mu \leq 1 - \exp(-D)$. Alternatively, using (3.5), $\mathcal{E}$ is dimension D-optimal if and only if $-\log(\exp(D) - 1) \leq \overline{D}_\mathcal{E}(u) \leq \log(\exp(D) - 1)$.

If $D \geq \log(3)$, then the dimension D-optimal exponential families are zero-dimensional, consisting of a single point $\{\nu\}$ such that $\min\{\nu_1, \nu_2, \nu_3\} \geq e^{-D}$. There are also one-dimensional inclusion D-optimal exponential families: Consider, for example, the exponential family $\mathcal{E}$ with sufficient statistics $A = (0, 1, 2)$ and reference measure $\nu = (1, 4, 1)$. The two local maximizers are $u^+ = \delta_2$ and $u^- = \frac{1}{2}(\delta_1 + \delta_3)$. Their rI-projection is $P_\mathcal{E} = \frac{1}{6}\nu$. Hence $D_\mathcal{E}(u^+) = \log\frac{3}{2}$ and $D_\mathcal{E}(u^-) = \log 3$, and so $\max D_\mathcal{E} =$

$\log 3$. The monomial parametrization of $\mathcal{E}$ is

$$P_\xi = \frac{1}{Z_\xi}(1, 4\xi, \xi^2),$$

where $\xi \in \mathbb{R}_\geq$ and $Z_\xi = 1 + 4\xi + \xi^2$. Consequently, $\mathcal{E}$ does not contain the uniform distribution. Therefore, any point $P \in \mathcal{E}$ satisfies $\max D(\cdot\|P) \geq \max D_\mathcal{E}$.

The following theorem generalizes the special case of Theorem 4.19 when N is even.

Theorem 5.5. *Let $\mathcal{X}$ be a finite set of cardinality N. Then $D_{N,k} \geq \log(N/(k+1))$ for all $0 \leq k < N$. If $\mathcal{E}$ is a k-dimensional exponential family that satisfies $\max D_\mathcal{E} = \log(N/(k+1))$, then $\mathcal{E}$ is a partition model of a homogeneous partition of coarseness $N/(k+1)$. In particular, if N is divisible by $(k+1)$, then $D_{N,k} = \log(N/(k+1))$, and the dimension $D_{N,k}$-optimal models are partition models.*

Proof. First assume that $\mathcal{E} \in \mathcal{H}_\mathbf{1}$. Let A be a sufficient statistics of $\mathcal{E}$. The moment map π_A maps the uniform distribution $Q = \frac{1}{N}\mathbf{1}$ to a point in the interior of $\mathbf{M}_A$. By Carathéodory's theorem (Theorem A.4) there are $k+1$ vertices $A_{x_0}, \ldots, A_{x_k}$ of $\mathbf{M}_A$ and $\lambda_0, \ldots, \lambda_k \in \mathbb{R}_\geq$ such that $\pi_A(Q) = \sum_{i=0}^k \lambda_i A_{x_i}$ and $\sum_{i=0}^k \lambda_i = 1$. Let $P = \sum_{i=0}^k \lambda_i \delta_{x_i}$, then $Q = P_\mathcal{E}$. By the Pythagorean theorem, $\max D_\mathcal{E} \geq D_\mathcal{E}(P) = H(Q) - H(P) \geq \log(N) - \log(k+1)$, proving the first assertion.

If equality holds, then $\lambda_0 = \cdots = \lambda_k = \frac{1}{k+1}$. Let $x \in \mathcal{X} \setminus \{x_0, \ldots, x_k\}$. For $i \in \{0, \ldots, k\}$ let C_i be the convex hull of $A_{x_0}, \ldots, A_{x_{i-1}}, A_{x_{i-1}}, \ldots, A_{x_k}$ and A_x. By Carathéodory's theorem the sets C_i cover the convex hull of $A_{x_0}, \ldots, A_{x_k}$ and A_x. In particular, $\pi_A(Q) \in C_j$ for some $j \in \{0, \ldots, k\}$, so $\pi_A(Q) = \sum_{i \neq j} \lambda'_i A_{x_i} + \lambda'_j A_x$. By the same argument as above it follows that $\lambda'_0 = \cdots = \lambda'_k = \frac{1}{k+1}$. Therefore, $A_x = (k+1)\pi_A(Q) - \sum_{i \neq j} A_{x_i} = A_{x_j}$.

Let $\sim$ be the equivalence relation on $\mathcal{X}$ defined by $x \sim y$ if and only if $A_x = A_y$, and let $\mathcal{X}' = (\mathcal{X}^1, \ldots, \mathcal{X}^{N'})$ be the corresponding partition into equivalence classes. Then $N' \leq k+1$ by what was shown until now. From $\dim(\mathcal{E}) = \dim(\mathbf{M}_A)$ one concludes $N' = k+1$, and $\mathbf{M}_A$ is a simplex of dimension k. By Lemma 4.10, $\mathcal{E}$ equals the partition model of $\mathcal{X}'$. Lemma 4.14 implies that the coarseness of $\mathcal{X}'$ equals $\frac{N}{k+1}$, which must be an integer. Furthermore, $\mathcal{X}'$ is homogeneous.

It remains to prove $\max D_\mathcal{E} > \log(N/(k+1))$ in the case $\mathcal{E} \notin \mathcal{H}_\mathbf{1}$. Let $\mathcal{N}_\mathbf{1}$ be the fibre of the uniform distribution. The function $D_\mathcal{E}$ is convex on $\mathcal{N}_\mathbf{1}$, hence $D_\mathcal{E}$ is maximal at the vertices of $\mathcal{N}_\mathbf{1}$, and any such vertex P satisfies $|\operatorname{supp}(P)| \leq \dim(\mathcal{X})+1$ (cf. the proof of Lemma 3.5). Let $P_\mathcal{E}$ be the rI-projection of the uniform distribution. Denote by $\mathcal{E}_\mathbf{1}$ the exponential family with uniform reference measure and with the same normal space as $\mathcal{E}$. On $\mathcal{N}_\mathbf{1}$ the difference

$$\delta(P) := D_\mathcal{E}(P) - D_{\mathcal{E}_\mathbf{1}}(P) = -\sum_{x \in \mathcal{X}} P(x) \log P_\mathcal{E}(x) - \log N$$

is an affine function that is positive at the uniform distribution. Hence there is a vertex P of $\mathcal{N}_\mathbf{1}$ such that $\delta(P) > 0$, and so $D_\mathcal{E}(P) > D_{\mathcal{E}_\mathbf{1}}(P) = \log N - H(P) \geq \log(N/(k+1))$. □

The value of $D_{N,k}$ is unknown when $k+1$ does not divide N. The situation is known for $N = 3$, see Example 5.4: If $1 \leq k < 3$, then $D_{N,k} = \log(2)$, and all dimension $D_{N,1}$-optimal exponential families that contain the uniform distribution are partition models. The following conjecture generalizes this example and Theorems 4.19 and 5.5:

Conjecture 5.6. $D_{N,k} = \log\lceil \frac{N}{k+1} \rceil$, and the dimension $D_{N,k}$-optimal exponential families containing the uniform distribution are partition models.

The following weaker statement holds:

Lemma 5.7. *Let $\mathcal{X}' = \{\mathcal{X}^1, \ldots, \mathcal{X}^{N'}\}$ be a partition of coarseness $c < N$ such that $\mathcal{X}^1$ has cardinality $l \leq c$ and all other components $\mathcal{X}^i$ for $i > 1$ have cardinality c. Then the partition model $\mathcal{E}$ of $\mathcal{X}'$ is $\log(c)$-inclusion optimal.*

Proof. The fact that $\max D_{\mathcal{E}} = \log(c)$ follows from Lemma 4.14. It remains to prove the optimality. Let $\mathcal{E}' \subseteq \mathcal{E}$ be an exponential family contained in $\mathcal{E}$. Let $\mathcal{Z}$ be the union of all blocks of $\mathcal{X}'$ of cardinality c. Assume that there exists a probability measure $Q \in \overline{\mathcal{E}} \setminus \overline{\mathcal{E}'}$ with support contained in $\mathcal{Z}$. By Corollary 4.15 there exists $P \in \mathbf{P}(\mathcal{Z})$ such that $Q = P_{\mathcal{E}}$ and $D(P\|Q) = \log(c)$. Let $Q' = P_{\mathcal{E}'} \in \mathcal{E}$. Then $D(P\|Q') = D(P\|Q) + D(Q\|Q') > \log(c)$ by the Pythagorean identity. Otherwise, if $\overline{\mathcal{E}} \cap \mathbf{P}(\mathcal{Z}) = \overline{\mathcal{E}'} \cap \mathbf{P}(\mathcal{Z})$, then $\dim(\mathcal{E}) = \dim(\overline{\mathcal{E}} \cap \mathbf{P}(\mathcal{Y})) + 1 = \dim(\overline{\mathcal{E}'} \cap \mathbf{P}(\mathcal{Y})) + 1 \leq \dim(\mathcal{E}')$, so $\mathcal{E} = \mathcal{E}'$. □

Theorem 5.5 can be applied to the hierarchical models $\mathcal{E}_K$ for $K \subseteq [n]$ introduced in Remark 4.11. By Theorem 5.5 the hierarchical model $\mathcal{E}_K$ is dimension optimal with $\max D(\cdot\|\mathcal{E}_K) = \sum_{i \in [n]\setminus K} \log(N_i)$. If $N_n = 2$, then the choice $K = \{1, \ldots, n-1\}$ yields an exponential family of dimension less than $|\mathcal{X}|/2$ such that $\max D(\cdot\|\mathcal{E}_K) = \log(2)$, and Theorem 4.18 implies that $\mathcal{E}_K$ is dimension optimal. The following proposition says that the exponential families $\mathcal{E}_K$ are the unique dimension D-optimal hierarchical models for many values of D.

Proposition 5.8. *Let $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$, where $N_i = |\mathcal{X}_i| < \infty$. For any $K \subseteq [n]$ let $D_K = \sum_{i \notin K} \log(N_i)$. The hierarchical model $\mathcal{E}_K$ is dimension D_K-optimal.*

Let l be any divisor of $N := |\mathcal{X}| = \prod_{i=1}^n N_i$. If $\mathcal{E}$ is any hierarchical model that is dimension $\log(N/l)$-optimal, then there is a subset $K \subseteq [n]$ such that $\mathcal{E} = \mathcal{E}_K$.

The proposition implies that if l is not of the form $\prod_{i \in K} N_i$ for some subset $K \subseteq [n]$, then there exists no hierarchical model that is dimension $\log(N/l)$-optimal.

Proof. It only remains to prove the last statement. If $\mathcal{E}$ satisfies the assumptions, then $\mathcal{E}$ is a partition model by Theorem 5.5. Therefore, it suffices to prove that any hierarchical model that is also a partition model is of the form $\overline{\mathcal{E}_K}$.

Let Δ be a simplicial complex on $[n]$ such that $\mathcal{E} = \mathcal{E}_\Delta$, and let $K = \cup_{J \in \Delta} J$. Then $\mathcal{E}$ is a submodel of $\mathcal{E}_K$. Let A be a sufficient statistics of $\mathcal{E}$. By Lemma 2.45 the convex supports of $\mathcal{E}$ and $\mathcal{E}_K$ have the same number of vertices. By Lemma 4.10 they are both simplices, hence they have the same dimension, so $\mathcal{E} = \mathcal{E}_K$. □

Conjecture 5.6 would imply that the partition models of Lemma 5.7 are dimension optimal among all exponential families. If the conjecture were true, then it would suggest the following interpretation: In many cases the information divergence $D(P\|Q)$ can be interpreted as the information which is lost when P is the true probability distribution, but computations are carried out with Q. For example, in the case of the independence model $\mathcal{E}_1$ of two variables, $D_{\mathcal{E}_1}$ equals the mutual information and measures the amount of information that one variable carries about the other variable. If a probability measure is replaced by its rI-projection, then this information is lost.

For the exponential families $\mathcal{E}_K$ the loss equals $D_K = \sum_{i \notin K} \log(N_i)$, which is precisely the maximal information that the random variables that are not in K can carry. Assuming that the conjecture is true, if the model is smaller than $\mathcal{E}_K$, then, in general, more information can be lost. In this interpretation the fact that $\max D_{\mathcal{E}} \geq \log(2)$ unless $\mathcal{E} = \mathbf{P}(\mathcal{X})^\circ$ means that for any exponential family $\mathcal{E} \neq \mathbf{P}(\mathcal{X})^\circ$ in general at least one bit is necessary to compensate the approximation of arbitrary probability measures.

The results of this section suggest the following strategy for the minimax algorithm: Start with $F^0 = \emptyset$. At each step in the algorithm choose $\mathcal{F}^k$ such that for all $F \in \mathcal{F}^k$ the exponential family $\mathcal{E}_{F^k \cup F}$ is a partition exponential family of some partition. Let $\mathcal{X}'_k$ and $\mathcal{X}'_{k,F}$ be the partitions of $\mathcal{E}_{F^k}$ and $\mathcal{E}_{F^k \cup F}$, respectively. Then the partitions $\mathcal{X}'_{k,F}$ refine the partition $\mathcal{X}'_k$. There are different possibilities to make these ideas concrete:

- One may require that the partitions are homogeneous partitions. In particular, it is possible to restrict to homogeneous partitions of coarseness $2^{|\mathcal{X}|-k}$.
- Alternatively, one may choose the candidate features such that at each step one (or more) of the largest blocks in the partition is split.

The idea of refining the blocks of the partition one by one has the advantage that $\mathcal{F}^k$ can be chosen such that any $F \in \mathcal{F}^k$ consists of only one feature f and such that this feature $f \in \{0,1\}^{\mathcal{X}}$ takes only two values. In [24] such features have been called *binary features*, and it was shown that such features can be treated efficiently. It would be interesting to apply these ideas at an example. This will be done in a future project.

5.3. Asymptotic behaviour of the empirical information divergence

This section studies the asymptotic distribution of the empirical information divergence from an exponential family, generalizing Milan Studený's results on the empirical multiinformation [68]. For the independence models the rI-projection map $P \mapsto P_{\mathcal{E}}$ is known in closed form. This makes it possible to make the results for the case when the distribution P is a projection point more explicit. Studený's results can be derived from Theorem 5.11 below, as will be shown in Remark 5.16.

This section is the only part of this thesis in which infinite probability spaces occur: The discussion of the asymptotic behaviour of estimators makes it necessary to study

infinite sequences of independent and identically distributed random variables. There are no principal difficulties to make the following treatment formally precise in the setting of Kolmogorov's axioms. Some of the technical details are discussed in [68].

For any probability measure P on $\mathcal{X}$ let

$$R[P] := \sum_{x \in \mathrm{supp}(P)} P(x) \log^2 \left(\frac{P(x)}{P_{\mathcal{E}}(x)} \right) - D(P\|\mathcal{E})^2$$

be the variance of the random variable $x \mapsto \log \left(\frac{P(x)}{P_{\mathcal{E}}(x)} \right)$ under P. Then $R[P] \geq 0$, and $R[P] = 0$ if and only if $\frac{P(x)}{P_{\mathcal{E}}(x)}$ is a constant, which means that P is a projection point.

Definition 5.9. Let $(X^{(i)})_{i \geq 1}$ be a sequence of independent random variables, identically distributed according to P. For any $n > 0$ let

$$\hat{P}^{(n)}(x) = \frac{1}{n} \left| \left\{ i \leq n : X^{(i)} = x \right\} \right| \tag{5.2}$$

be the *empirical distribution* after n steps. Every empirical distribution is a random variable with values in $\mathbf{P}(\mathcal{X})$. For any n the *empirical information divergence* is the random variable

$$\hat{D}_{\mathcal{E}}^{(n)} := D(\hat{P}^{(n)}\|\mathcal{E}). \tag{5.3}$$

Definition 5.10. The *normal distribution* on $\mathbb{R}^n$ with mean $b \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$ is denoted by $N(b, \Sigma)$.

Let $Y^1, \ldots, Y^n, \ldots$ be a sequence of random variables, and let P be a probability measure. If P is not a point measure, then Y^n is *asymptotically P-distributed* if there exist sequences $(a_n)_n, (b_n)_n \subset \mathbb{R}$ such that $a_n > 0$ for all n and such that $\frac{1}{a_n}(Y^n - b_n)$ converges in distribution to a P-distributed random variable. If $P = N(0,1)$, then Y^n is *asymptotically $N(b_n, a_n^2)$-distributed*.

Let A be a sufficient statistics of the exponential family $\mathcal{E}$ such that the rows of A are linearly independent and such that $\mathbf{1}$ is not contained in the row space of A. For any positive measure $\mu \in \mathbb{R}^{\mathcal{X}}$ let $\mu_{\mathcal{E}}$ be the rI-projection of $\frac{1}{\mu(\mathcal{X})}\mu$ onto $\mathcal{E}$, and let $m_i(\mu) = \sum_x A_{i,x}\mu_{\mathcal{E}}(x) = \frac{1}{\mu(\mathcal{X})} \sum_x A_{i,x}\mu_x$. Let $V(\mu)^{-1}$ be the inverse of the covariance matrix

$$V(\mu)_{i,j} = \sum_x \mu_{\mathcal{E}}(x) A_{i,x} A_{j,x} - \sum_x \mu_{\mathcal{E}}(x) A_{i,x} \sum_y \mu_{\mathcal{E}}(y) A_{j,x}$$

of A at $\mu_{\mathcal{E}}$. This inverse exists by the assumptions on A: If $\sum_j V_{i,j}\vartheta_j = 0$ for some $\vartheta \in \mathbb{R}^h$, then the function $u = \sum_j \vartheta_j A_{j,x} \in \mathbb{R}^{\mathcal{X}}$ has variance $\sum_{i,j} \vartheta_i V_{i,j} \vartheta_j = 0$. Therefore, u is a constant function in the row span of A. By assumption, $u = 0$, and hence $\vartheta = 0$, since A is invertible.

Theorem 5.11. *Let $P \in \mathbf{P}(\mathcal{X})$. If P is not a projection point, then the empirical information divergence is asymptotically $N(D(P\|\mathcal{E}), \frac{1}{n}R[P])$-distributed. Otherwise,*

let $\lambda_1, \ldots, \lambda_N$ be the eigenvalues of the $\mathcal{X} \times \mathcal{X}$-matrix with matrix elements

$$g(x,y) = \frac{1}{2}\left(\delta_{x,y} - \sum_{i,j} A_{i,x} V(P)^{-1}_{i,j} \left(A_{j,y} - m_j(P)\right) P(y)\right),$$

for all $x, y \in \mathcal{X}$. Order the eigenvalues such that $\lambda_N = \frac{1}{2}$. Then $n(\hat{D}^{(n)}_{\mathcal{E}} - D_{\mathcal{E}})$ converges in distribution to $\sum_{i=1}^{N-1} \lambda_i t_i^2$, where $t_1, \ldots, t_{N-1}$ are independent $N(0,1)$-distributed random variables.

Distributions of random variables of the form $\sum_{i=1}^{N-1} \lambda_i t_i^2$ for independent $N(0,1)$-distributed random variables t_i are sometimes called *generalized χ^2-distributions*. The proof of the theorem will be given after a series of preliminary lemmas. The idea of the proof is to do a Taylor approximation of $D_{\mathcal{E}}$. The first lemma computes the necessary partial derivatives of $D_{\mathcal{E}}$.

Lemma 5.12. *Let $D_{\mathcal{E}}(\mu) = D(\mu\|\mathcal{E})$ for any positive measure $\mu \in \mathbb{R}^{\mathcal{X}}$. Then*

$$\frac{\partial}{\partial \mu_x} D_{\mathcal{E}}(\mu) = 1 + \log \frac{\mu_x}{\mu_{\mathcal{E}}(x)},$$
$$\frac{\partial^2}{\partial \mu_x \partial \mu_y} D_{\mathcal{E}}(\mu) = \frac{\delta_{x,y}}{\mu_x} - \frac{1}{\mu(\mathcal{X})} \sum_{i,j} (A_{i,x} - m_i(\mu)) V(\mu)^{-1}_{i,j} (A_{j,y} - m_j(\mu)),$$
for all $x, y \in \operatorname{supp}(\mu)$.

Proof. Let $\mathcal{Z} = \operatorname{supp}(\mu)$. From Proposition 2.14 (i) it follows that

$$D_{\mathcal{E}}(a\nu) = a\left(D_{\mathcal{E}}(\nu) + \nu(\mathcal{X}) \log(a)\right), \qquad \text{for all } a \in \mathbb{R}. \tag{5.4}$$

Fix $x_0 \in \mathcal{Z}$. Consider the set of coordinates $\{\overline{\mu}, \mu_x\}_{x \in \mathcal{Z} \setminus \{x_0\}}$, where $\overline{\mu} := \mu(\mathcal{X})$. In order to distinguish the partial derivatives with respect to this coordinate system from the partial derivatives with respect to the coordinate system $\{\mu_x\}_{x \in \mathcal{Z}}$ the notation $\left(\frac{\partial}{\partial \mu_x} D_{\mathcal{E}}\right)_{\overline{\mu}}$ and $\left(\frac{\partial}{\partial \mu_x} D_{\mathcal{E}}\right)_{\mu_{x_0}}$ from statistical thermodynamics is used. The two coordinate systems are related via $\mu_{x_0} = \overline{\mu} - \sum_{x \neq x_0} \mu_x$. Write $\mu = \overline{\mu} Q$, where Q is a probability measure. By Theorem 3.1,

$$\left(\frac{\partial}{\partial \mu_x} D_{\mathcal{E}}(Q)\right)_{\overline{\mu}} = \frac{1}{\overline{\mu}} \left(\log \frac{Q(x)}{\mu_{\mathcal{E}}(x)} - \log \frac{Q(x_0)}{\mu_{\mathcal{E}}(x_0)}\right) = \frac{1}{\overline{\mu}} \left(\log \frac{\mu_x}{\mu_{\mathcal{E}}(x)} - \log \frac{\mu_{x_0}}{\mu_{\mathcal{E}}(x_0)}\right),$$

for all $x \in \mathcal{Z} \setminus \{x_0\}$. Therefore,

$$\left(\frac{\partial}{\partial \mu_x} D_{\mathcal{E}}(\mu)\right)_{\overline{\mu}} = \frac{\partial}{\partial \mu_x} \left(\overline{\mu}\left(D_{\mathcal{E}}(Q) + \log(\overline{\mu})\right)\right) = \log \frac{\mu_x}{\mu_{\mathcal{E}}(x)} - \log \frac{\mu_{x_0}}{\mu_{\mathcal{E}}(x_0)}$$

for all $x \in \mathcal{Z} \setminus \{x_0\}$. Furthermore, using (5.4),

$$\nu(\mathcal{X})\frac{\partial}{\partial\overline{\mu}}D_{\mathcal{E}}(a\nu) + \sum_{x\in\mathcal{Z}\setminus\{x_0\}} \nu_x \left(\frac{\partial}{\partial\mu_x}D_{\mathcal{E}}(a\nu)\right)_{\overline{\mu}} = \frac{\partial}{\partial a}D_{\mathcal{E}}(a\nu)$$
$$= D_{\mathcal{E}}(\nu) + \nu(\mathcal{X})(\log(a)+1) = \frac{1}{a}D_{\mathcal{E}}(a\nu) + \nu(\mathcal{X}).$$

For $a = \overline{\mu}$ and $\nu = Q$ this yields

$$\begin{aligned}\frac{\partial}{\partial\overline{\mu}}D_{\mathcal{E}}(\mu) &= \frac{1}{\overline{\mu}}D_{\mathcal{E}}(\mu) + 1 - \sum_{x\in\mathcal{Z}\setminus\{x_0\}} Q(x)\left(\frac{\partial}{\partial\mu_x}D_{\mathcal{E}}(\mu)\right)_{\overline{\mu}} \\ &= \frac{1}{\overline{\mu}}D_{\mathcal{E}}(\mu) + 1 - \sum_{x\in\mathcal{Z}} Q(x)\left(\log\frac{\mu_x}{\mu_{\mathcal{E}}(x)} - \log\frac{\mu_{x_0}}{\mu_{\mathcal{E}}(x_0)}\right) \\ &= 1 + \log\frac{\mu_{x_0}}{\mu_{\mathcal{E}}(x_0)}\end{aligned}$$

Hence

$$\left(\frac{\partial}{\partial\mu_y}D_{\mathcal{E}}(\mu)\right)_{\mu_{x_0}} = \frac{\partial}{\partial\overline{\mu}}D_{\mathcal{E}}(\mu) + \left(\frac{\partial}{\partial\mu_y}D_{\mathcal{E}}(\mu)\right)_{\overline{\mu}} = 1 + \log\frac{\mu_y}{\mu_{\mathcal{E}}(y)},$$
$$\frac{\partial}{\partial\mu_{x_0}}D_{\mathcal{E}}(\mu) = \frac{\partial}{\partial\overline{\mu}}D_{\mathcal{E}}(\mu) \qquad = 1 + \log\frac{\mu_{x_0}}{\mu_{\mathcal{E}}(x_0)}.$$

This proves the first formula.

Deriving the first formula with respect to μ_y yields

$$\frac{\partial^2}{\partial\mu_x\partial\mu_y}D_{\mathcal{E}}(\mu) = \frac{\delta_{x,y}}{\mu_x} - \frac{1}{\mu_{\mathcal{E}}(\mathcal{X})}\frac{\partial\mu_{\mathcal{E}}(x)}{\partial\mu_y}.$$

The map $\mu \mapsto \mu_{\mathcal{E}}$ factors as the composition of the map $m : \mu \mapsto \frac{1}{\mu(\mathcal{X})}A\mu$, the inverse of $\vartheta \mapsto AP_\vartheta$ and $\vartheta \mapsto P_\vartheta$. Note that the second map $\vartheta \mapsto AP_\vartheta$ is itself the composition of the third map $\vartheta \mapsto P_\vartheta$ and m. The differential of the first map is

$$\frac{\partial m(\mu)}{\partial\mu_x} = \frac{1}{\mu(\mathcal{X})}\left(A_x - m(\mu)\right).$$

The differential of the third map is

$$\frac{\partial}{\partial\vartheta_i}P_\vartheta(x) = (A_{i,x} - m_i(P_\vartheta))P_\vartheta(x).$$

Hence, by the chain rule, the differential of the second map is

$$\frac{\partial}{\partial\vartheta_j}(m_i(P_\vartheta)(x)) = \sum_x \left(A_{i,x} - m_i(P_\vartheta)\right)\left(A_{j,x} - m_j(P_\vartheta)\right)P_\vartheta(x) = V(P_\vartheta)_{i,j}.$$

Now the second formula follows from $m(\mu_{\mathcal{E}}) = m(\mu)$ and

$$\frac{\partial\mu_{\mathcal{E}}(x)}{\partial\mu_y} = \sum_{i,j}(A_{i,x} - m_i(\mu))\mu_{\mathcal{E}}(x)V(\mu)^{-1}_{i,j}\frac{1}{\mu(\mathcal{X})}(A_{j,y} - m_j(\mu)). \qquad \square$$

The next three lemmas are needed to compute the asymptotic distribution of the different terms in the Taylor expansion:

Lemma 5.13. *Let $C = (c(x,y))_{x,y\in\mathcal{X}} \in \mathbb{R}^{\mathcal{X}\times\mathcal{X}}$ be the matrix with matrix elements $c(x,y) = P(x)(\delta_{x,y} - P(y))$. Then $\hat{P}^{(n)}$ has expectation P and covariance matrix $\frac{1}{n}C$. The normalized empirical probability $\delta\hat{P}^{(n)} := \sqrt{n}\left(\hat{P}^{(n)} - P\right)$ has expectation zero and covariance C and is asymptotically $N(0,C)$ distributed.*

Proof. See [68, Proposition 2]. □

Lemma 5.14. *Let $B, C \in \mathbb{R}^{N\times N}$ be two symmetric real matrices, and assume that C is positive definite. Then BC is diagonalizable, and all eigenvalues $\lambda_1, \ldots, \lambda_N$ of BC are real. If X is a $N(0,C)$-distributed random vector, then the random variable $\sum_{i,j\in\mathcal{X}} B_{i,j}X_iX_j$ has the same distribution as $\sum_{i=1}^N \lambda_i t_i^2$, where $t_1, \ldots, t_N$ are independent $N(0,1)$-distributed random variables.*

Proof. See [68, Lemma 4]. □

See the proof of Theorem 5.11 how to apply the last lemma in the case that C is only positive semidefinite.

Lemma 5.15. *Let $G = (g_{i,j}) \in \mathbb{R}^{N\times N}$ be a diagonalizable matrix with eigenvalues $\lambda_1, \ldots, \lambda_N$. Assume that $\sum_j g_{i,j} = \lambda_N \in \mathbb{R}$. Then the eigenvalues of the matrix $D = (d_{i,j}) \in \mathbb{R}^{(N-1)\times(N-1)}$ defined by $d_{i,j} = g_{i,j} - g_{Z,j}$ are $\lambda_1, \ldots, \lambda_{N-1}$.*

Proof. Under the assumptions of the lemma the proof of [68, Lemma 5] applies. □

Proof of Theorem 5.11. Let $\mathcal{Z} = \operatorname{supp}(P)$. Consider $D_\mathcal{E} : \mu \mapsto D(\mu\|\mathcal{E})$ as a function from $(0,\infty)^\mathcal{Z}$ to $\mathbb{R}$. By Lemma 5.12, the Taylor expansion of $D_\mathcal{E}$ around P is

$$D_\mathcal{E}(P+u) = D_\mathcal{E}(P) + \sum_{x\in\mathcal{X}} u(x)\left(1 + \log\frac{P(x)}{P_\mathcal{E}(x)}\right) + \frac{1}{2}\sum_{x,y\in\mathcal{X}} u(x)u(y)\left(\frac{\delta_{x,y}}{P(x)} - \sum_{i,j}(A_{i,x} - m_i(P))V(P)^{-1}_{i,j}(A_{j,y} - m_j(P))\right) + \epsilon(u),$$

where $\epsilon(u) = o(u^3)$. Let $u^{(n)} = \hat{P}^{(n)} - P = \frac{1}{\sqrt{n}}\delta\hat{P}^{(n)}$. By Lemma 5.13 $u^{(n)}$ is asymptotically $N(0, \frac{1}{n}C)$-distributed. From

$$\sum_{x,y}\left(1 + \log\frac{P(x)}{P_\mathcal{E}(x)}\right)c(x,y)\left(1 + \log\frac{P(x)}{P_\mathcal{E}(x)}\right) = R[P]$$

and the general theory of asymptotic statistics (see, for example, Proposition 5.10 and Theorem 6a.2 (ii) in [59]) it follows that

$$\sqrt{\frac{n}{R[P]}}(\hat{D}^{(n)}_\mathcal{E} - D(P\|\mathcal{E})) \to N(0,1) \qquad \text{in distribution.}$$

Now assume that $R[P] = 0$. Then $\sum_{x\in\mathcal{X}} u^{(n)}(x)\left(1+\log\frac{P(x)}{P_{\mathcal{E}}(x)}\right) = 0$. Furthermore, Lemma 5.13 implies that the error term satisfies $n\epsilon(u^{(n)}) \to 0$ in probability. Therefore, it suffices to analyze the asymptotic behaviour of

$$\begin{aligned} k^{(n)} &:= \frac{n}{2}\sum_{x,y\in\mathcal{X}} u^{(n)}(x)u^{(n)}(y)\left(\frac{\delta_{x,y}}{P(x)} - \sum_{i,j}(A_{i,x}-m_i(P))V(P)^{-1}_{i,j}(A_{j,y}-m_j(P))\right) \\ &= \frac{1}{2}\sum_{x,y\in\mathcal{X}} \delta\hat{P}^{(n)}(x)\delta\hat{P}^{(n)}(y)\left(\frac{\delta_{x,y}}{P(x)} - \sum_{i,j} A_{i,x}V(P)^{-1}_{i,j}A_{j,y}\right), \end{aligned}$$

where the last equality follows from $\sum_{x\in\mathcal{X}} u^{(n)}(x) = 0$. Let

$$b'(x,y) = \frac{1}{2}\left(\frac{\delta_{x,y}}{P(x)} - \sum_{i,j} A_{i,x}V(P)^{-1}_{i,j}A_{j,y}\right)$$

Fix $z_0 \in \mathcal{X}$, and let $\mathcal{X}' = \mathcal{X}\setminus\{z_0\}$. Then $\delta\hat{P}^{(n)}(z_0) = -\sum_{x\in\mathcal{X}'}\delta\hat{P}^{(n)}(x)$. Therefore, $k^{(n)}(x) = \sum_{x,y\in\mathcal{X}'} b(x,y)\delta\hat{P}^{(n)}(x)\delta\hat{P}^{(n)}(y)$, where $b(x,y) = b'(x,y) - b'(x,z_0) - b'(z_0,x)+b'(z_0,z_0)$. Let $\lambda_1,\dots,\lambda_{N-1}$ be the eigenvalues of the matrix $D \in \mathbb{R}^{\mathcal{X}'\times\mathcal{X}'}$ with matrix elements $d(x,y) = \sum_{z\in\mathcal{X}} b(x,z)c(z,y)$, where $c(z,y)$ is defined in Lemma 5.13. The restriction $C_{\mathcal{X}'\times\mathcal{X}'} \in \mathbb{R}^{\mathcal{X}'\times\mathcal{X}'}$ of C has full rank, because $\sum_{x\in\mathcal{X}}\hat{P}^{(n)}(x) = 1$ is the only relation between the components $\hat{P}^{(n)}(x)$ of the random vector $\hat{P}^{(n)}$ (alternatively, note that $C_{\mathcal{X}'\times\mathcal{X}'}$ has the same structure as $V(P)_{i,j}$ in the case of an independence model, see Remark 5.16; therefore an inverse can be found using (5.5)). Therefore, Lemma 5.14 applies and describes the distribution of $k^{(n)}$. It remains to characterize the spectrum of $BC_{\mathcal{X}'\times\mathcal{X}'}$.

Note that $\sum_{x\in\mathcal{X}} c(x,y) = 0$ for all $y \in \mathcal{X}$. Hence

$$\sum_{z\in\mathcal{X}'} b(x,z)c(z,y) = \sum_{z\in\mathcal{X}} b'(x,z)c(z,y) - \sum_{z\in\mathcal{X}} b'(z_0,z)c(z,y),$$

for all $x,y \in \mathcal{X}'$. Let $x \in \mathcal{X}$ and $y \in \mathcal{X}'$. Then

$$\begin{aligned} 2\sum_{z\in\mathcal{X}} b'(x,z)c(z,y) &= \frac{P(x)(\delta_{x,y}-P(y))}{P(x)} - \sum_{i,j} A_{i,x}V(P)^{-1}_{i,j}\sum_{z\in\mathcal{X}} A_{j,z}P(z)(\delta_{z,y}-P(y)) \\ &= \delta_{x,y} - P(y) - \sum_{i,j} A_{i,x}V(P)^{-1}_{i,j}\left(A_{j,y}-m_j(P)\right)P(y) \\ &= 2g(x,y) - P(y), \end{aligned}$$

so $\sum_{z\in\mathcal{X}'} b(x,z)c(z,y) = g(x,y) - g(z_0,y)$ for all $x,y \in \mathcal{X}'$. Note that $\sum_{z\in\mathcal{X}} g(x,z) = \frac{1}{2}$. Therefore, the statement of the theorem follows from Lemma 5.15. □

Remark 5.16. The remainder of this section discusses how to obtain Studený's original result from Theorem 5.11. Let $\mathcal{E} = \mathcal{E}_1$ be the independence model of n random

variables $X_1, \ldots, X_n$, where X_i takes values in $\mathcal{X}_i = [N_i]$ (see Section 2.4 for the notation). For any $i \in [n]$ let $x|_i = \pi_{\{i\}}(x)$ for all $x \in \mathcal{X}$, and write $P_i(x_i) = \sum_{x \in \mathcal{X}: x|_i = x_i} P(x)$. A sufficient statistics A satisfying the assumptions of Lemma 5.12 is given by

$$A_{(i,x_i),x} = \delta_{x_i,x|_i}, \qquad \text{for } i = [n], 1 \leq x_i < N_i, x \in \mathcal{X}.$$

Then $P_{\mathcal{E}}(x) = \prod_{i=1}^n P_i(x|_i)$, so

$$V(P)_{(i,x_i),(j,y_j)} = \delta_{i,j} P_i(x_i) \left(\delta_{x_i y_j} - P_j(x_j)\right).$$

The inverse of $V(P)$ satisfies

$$V(P)^{-1}_{(i,x_i),(j,y_j)} = \delta_{i,j} \left(\frac{1}{P_i(N_i)} + \delta_{x_i y_j} \frac{1}{P_i(x_i)} \right). \tag{5.5}$$

Note that $m_{(i,x_i)}(P) = P_i(x_i)$. Hence

$$\begin{aligned} 2g(x,y) &= \delta_{x,y} - \sum_{i, x_i < N_i, y_i < N_i} \delta_{x_i,x|_i} \left(\frac{1}{P_i(N_i)} + \delta_{x_i y_i} \frac{1}{P_i(x_i)} \right) \left(\delta_{y_i,y|_i} - P_i(y_i)\right) P(y) \\ &= \delta_{x,y} + \sum_i P(y) \left(\frac{\delta_{N_i,y|_i}}{P_i(N_i)} - \frac{\delta_{x|_i y|_i}}{P_i(x|_i)} \right) \end{aligned}$$

This matrix $G = (g(x,y))_{x,y \in \mathcal{X}}$ differs from the matrix $E = (e(x,y))_{x,y \in \mathcal{X}}$, defined in [68] by

$$2e(x,y) = \delta_{x,y} - P(y) \sum_i \frac{\delta_{x|_i,y|_i}}{P_i(x|_i)}.$$

The matrices G and E satisfy $g(x,y) - g(x,z) = e(x,y) - e(x,z)$ and $\sum_y e(x,y) = \frac{1}{2}(1-m)$. Hence Lemma 5.15 implies:

Lemma 5.17. *Let $\mathcal{E}$ be an independence exponential family. Let $\lambda_1, \ldots \lambda_{N-1}$ be as in Theorem 5.11. Then the eigenvalues of E are $\lambda_1, \ldots, \lambda_{N-1}, \frac{1}{2}(1-m)$.*

Proposition 4 in [68] follows from the second part of Theorem 5.11 and this lemma.

A. Polytopes and oriented matroids

A.1. Polytopes

This section summarizes the basic definitions and facts concerning polytopes and polyhedra. Two introductory textbooks are [35] and [75].

Let V be a finite-dimensional vector space over $\mathbb{R}$. A subset $\mathbf{B} \subseteq V$ is *convex* if $p, q \in \mathbf{B}$ implies $(1-s)p+sq \in \mathbf{B}$ for all $0 \leq s \leq 1$. The smallest convex set containing a given set $B \subseteq V$ is called the *convex hull* of B. A *polytope* is the convex hull of a finite set. A *0-1-polytope* is the convex hull of a subset of $\{0,1\}^N \subset \mathbb{R}^N$ for some N. A *closed half-space* of V is a set of the form $\{p : l(p) \leq c\}$, where $c \in \mathbb{R}$ and $l \neq 0$ is a non-zero linear form on V. An *open half-space* is the complement of a closed half-space. A *hyperplane* is the boundary of a half-space. Any closed convex set $\mathbf{B}$ is the intersection of all closed half-spaces containing $\mathbf{B}$. A finite intersection of closed half-spaces is called a *polyhedron*.

Theorem A.1. *Any affine image of a polyhedron is a polyhedron.*

Proof. See [35, Ex. 2.6.4]. □

Theorem A.2. *Any bounded polyhedron is a polytope. Conversely, any polytope is a bounded polyhedron.*

Proof. By definition, a polytope is a linear image of a simplex. Any simplex is a polytope as well as a bounded polyhedron, therefore the statement follows from Theorem A.1. □

Theorem A.3 (Strict separation)**.** *Let $\mathbf{A}, \mathbf{B} \subseteq V$ be two nonempty closed convex sets. If $\mathbf{A}$ is bounded and if $\mathbf{A} \cap \mathbf{B} = \emptyset$, then there exists a linear form l on V and $c \in \mathbb{R}$ such that $l(x) < c < l(y)$ for all $x \in \mathbf{A}$ and $y \in \mathbf{B}$.*

Proof. See [35, Theorem 2.2.1]. □

Let $\mathbf{B}$ be a convex set. If $\mathbf{B}$ is nonempty, then the *dimension* of $\mathbf{B}$ is the dimension of the vector space generated by $\mathbf{B} - p$, where $p \in \mathbf{B}$ is arbitrary. A *face* of $\mathbf{B}$ is the intersection of $\mathbf{B}$ with a set $\{p : l(p) = c\}$, where $c \in \mathbb{R}$ and l is a linear form on V, such that $\mathbf{B}$ is contained in $\{p : l(p) \leq c\}$. Note that $\mathbf{B}$ and $\emptyset$ are always faces of $\mathbf{B}$; all other faces are called *proper faces*. Each proper face of $\mathbf{B}$ is the intersection of $\mathbf{B}$ with some hyperplane. A face of dimension zero is a *vertex*, a face of dimension one is an *edge*, and a maximal proper face is a *facet*.

Let $\mathbf{B}$ be a polytope. Then $\mathbf{B}$ is the convex hull of its vertices. Any face $\mathbf{F}$ of $\mathbf{B}$ is a polytope itself. Any face of $\mathbf{F}$ is also a face of $\mathbf{B}$. The intersection of faces of $\mathbf{B}$ is again a face. Therefore, $\mathbf{F}$ is the convex hull of those vertices of $\mathbf{B}$ which are contained in $\mathbf{F}$. Any proper face $\mathbf{F}$ of $\mathbf{B}$ is the intersection of all facets of $\mathbf{B}$ containing $\mathbf{F}$.

Theorem A.4 (Carathéodory's theorem)**.** *Let B be a subset of some real vector space. If the convex hull of A has dimension d, then every x in the convex hull of B is expressible in the form*

$$x = \sum_{i=0}^{d} \lambda_i x_i, \qquad \text{where } x_i \in A, \lambda_i \geq 0 \text{ and } \sum_{i=0}^{d} \lambda_i = 1.$$

Proof. See [35, Theorem 2.3.5]. □

There are different equivalence relations on the set of polytopes. In this thesis, the following notion will be important: Two polytopes $\mathbf{B} \subset V$ and $\mathbf{C} \subset V'$ are *affinely equivalent* if there exist affine maps $\beta : V \mapsto V'$ and $\gamma : V' \mapsto V$ such that $\beta(\mathbf{B}) = \mathbf{C}$ and $\gamma(\mathbf{C}) = \mathbf{B}$.

A.2. Oriented matroids

This appendix collects basic facts about matroids and oriented matroids. Only representable (oriented) matroids are needed, but general (oriented) matroids are mentioned for comparison. The textbooks [55] and [13] contain overviews over the general theory. See [61] for a short introduction that discusses the relation between oriented matroids and exponential families. Oriented matroids are also a valuable combinatorial tool for the study of polytopes and point configurations, see [13, Chapter 9], a connection that will not be pursued further here.

Let $\mathcal{N}$ be a subgroup of the additive group $\mathbb{R}^{\mathcal{X}}$ (usually $\mathcal{N}$ will be a linear subspace, but in the algebraic case $\mathcal{N}$ may be a sublattice of $\mathbb{Z}^{\mathcal{X}}$). A nonempty subset $\mathcal{Z} \subseteq \mathcal{X}$ is called *dependent* if there exists a vector $v \in \mathcal{N}$ such that $\mathcal{Z} = \operatorname{supp}(v)$. An inclusion minimal dependent set is called a *circuit*. If $\mathcal{Z}$ is a circuit, then any $v \in \mathcal{N}$ with $\mathcal{Z} = \operatorname{supp}(v)$ is called a *circuit vector*. The minimality condition implies that a circuit determines its corresponding circuit vector up to a multiple, i.e if $n' \in \mathcal{N}$ satisfies $\operatorname{supp}(n') \subseteq \operatorname{supp}(n)$, then $n' = \lambda n$ for some $\lambda \in \mathbb{R}$. The sign vector $\operatorname{sgn}(v)$ of a circuit vector v is a *signed circuit*.

In the algebraic case, when $\mathcal{N}$ is a vector space spanned by $\mathcal{N}_{\mathbb{Z}} = \mathcal{N} \cap \mathbb{Z}^{\mathcal{X}}$, every circuit has integer circuit vectors. In other words, the circuit vectors of $\mathcal{N}_{\mathbb{Z}}$ are the integer circuit vectors of $\mathcal{N}$. It is convenient to restrict attention to those integer circuit vectors $c \in \mathcal{N}_{\mathbb{Z}}$ that are as short as possible by requiring that the largest common divisor of the components $\{c(x) : x \in \mathcal{X}\}$ equals one. Such circuit vectors are called *prime*. To every circuit there are exactly two prime circuit vectors, and they differ by a factor of -1.

Let $\mathcal{C}$ be the set of signed circuits of $\mathcal{N}$, and let C be the set of circuits of $\mathcal{N}$. The pair $(\mathcal{X}, \mathcal{C})$ is called the *(representable) oriented matroid* of $\mathcal{N}$, and the pair $(\mathcal{X}, C)$ is called the *(representable) matroid* of $\mathcal{N}$. In the algebraic case the oriented matroids of $\mathcal{N}$ and $\mathcal{N}_{\mathbb{Z}}$ agree.

There are related methods to construct representable oriented matroids: If A is a matrix, then the oriented matroid of A is the oriented matroid of its kernel. In this case, a signed circuit corresponds to a minimal linear relation between the columns of A. More generally, an oriented matroid can be associated to any vector configuration $\{A_x\}_{x\in\mathcal{X}}$ in $\mathbb{R}^h$: Just interpret the vectors A_x as columns of an $h \times \mathcal{X}$-matrix A.

This last construction explains how Matroids serve as a model of dependence structures: A subset $\mathcal{Z} \subset \mathcal{X}$ is called *dependent* if it contains a circuit, otherwise it is *independent*. In other word, circuits are the minimal dependent sets. The maximal independent subsets of $\mathcal{X}$ are called *bases*. The *rank* $r(\mathcal{Z})$ of $\mathcal{Z} \subseteq \mathcal{X}$ is the cardinality of the largest independent subset of $\mathcal{Z}$. If the oriented matroid comes from a matrix A, then $r(\mathcal{Z})$ agrees with the rank of the submatrix $A_{\mathcal{Z}}$ that consists of those columns A_x with $x \in \mathcal{Z}$.

A *circuit basis* of $\mathcal{N}$ is a subset of $\mathcal{N}$ containing precisely one circuit vector for every circuit. It is easy to see that a circuit basis of a vector space $\mathcal{N}$ spans $\mathcal{N}$, see Lemma A.6 below, but in general the circuit vectors are not linearly independent.

Addition of vectors corresponds to the *composition* $\circ$ of sign vectors, where $\circ$ is the associative operation defined by

$$(\sigma_i \circ \sigma_{i+1})_x = \begin{cases} (\sigma_i)_x, & \text{if } (\sigma_i)_x \neq 0, \\ (\sigma_{i+1})_x, & \text{else.} \end{cases}$$

Two vectors $u, v \in \mathbb{R}^{\mathcal{X}}$ are *sign-consistent* if $u(x) \neq 0 \neq v(x)$ implies $\operatorname{sgn}(u(x)) = \operatorname{sgn}(v(x))$ for all $x \in \mathcal{X}$. For more detailed proofs of the following two lemmas see [13].

Lemma A.5. *For every nonzero vector $u \in \mathcal{N}$ there exists a sign-consistent circuit vector $c \in \mathcal{N}$ such that* $\operatorname{supp}(c) \subseteq \operatorname{supp}(u)$.

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

Lemma A.6. *Every $u \in \mathcal{N}$ is a finite sign-consistent sum $u = \sum_{i=1}^{r} c_i$ of circuit vectors $c_1, \ldots, c_r$, i.e. for all $x \in \mathcal{X}$ and for all i, if $c_i(x) \neq 0$, then* $\operatorname{sgn} c_i(x) = \operatorname{sgn} u(x)$.

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

The *dual (oriented) matroid* of the (oriented) matroid of a vector space $\mathcal{N} \subseteq \mathbb{R}^{\mathcal{X}}$ is the (oriented) matroid of the orthogonal complement $\mathcal{N}^{\perp}$. Clearly the dual oriented matroid of the dual oriented matroid of $\mathcal{N}$ equals the oriented matroid of $\mathcal{N}$. The (signed) circuits of the dual oriented matroid are called *(signed) cocircuits*. The rank

$r^*(\mathcal{Z})$ of a set $\mathcal{Z} \subseteq \mathcal{X}$ with respect to the dual matroid is also called the *corank* of $\mathcal{Z}$. If the matroid comes from a vector space $\mathcal{N}$, then let C be a matrix such that the rows of C span $\mathcal{N}$. Then $r^*(\mathcal{Z})$ agrees with the rank of the submatrix $C_{\mathcal{Z}}$ that consists of those columns C_x of C with $x \in \mathcal{Z}$.

Two sign vectors $\sigma, \tau \in \{0, \pm 1\}^{\mathcal{X}}$ are called *orthogonal* if the following holds: Either $\operatorname{supp}(\sigma) \cap \operatorname{supp}(\tau) = \emptyset$, or there exist $x, y \in \operatorname{supp}(\sigma) \cap \operatorname{supp}(\tau)$ such that $\sigma(x) = \tau(x)$ and $\sigma(y) \neq \tau(y)$. The fact that σ and τ are orthogonal is expressed by $\sigma \perp \tau$. By construction, if σ is a signed circuit and τ is a signed cocircuit, then $\sigma \perp \tau$.

Proposition A.7. *A sign vector $\sigma \in \{0, \pm 1\}^{\mathcal{X}}$ belongs to $\mathcal{N}$ if and only if $\sigma \perp \tau$ for all signed cocircuits τ of $\mathcal{N}$.*

Proof. See Proposition 3.7.12 in [13] □

Lemma A.5 proves that the representable oriented matroid of $\mathcal{N}$ is an oriented matroid in the following abstract sense:

Definition A.8. Let C be a set of subsets of $\mathcal{X}$. Then $(\mathcal{X}, C)$ is a *matroid* if

(C1) $\emptyset \notin C$,

(C2) for all $\mathcal{Y}, \mathcal{Z} \in C$, if $\mathcal{Y} \subseteq \mathcal{Z}$, then $\mathcal{Y} = \mathcal{Z}$,

(C3) for all $\mathcal{Y}, \mathcal{Y}' \in C$ and $e \in \mathcal{Y} \cap \mathcal{Y}'$ there exists $\mathcal{Z} \in C$ such that $\mathcal{Z} \subseteq (\mathcal{Y} \cup \mathcal{Y}') \setminus \{e\}$. (*weak elimination*)

Let $\mathcal{C} \subset \{0, \pm 1\}^{\mathcal{X}}$ be a set of sign vectors. For any $\sigma \in \mathcal{C}$ write $\underline{\sigma} = \operatorname{supp}(\sigma)$. Then $(\mathcal{X}, \mathcal{C})$ is an *oriented matroid* if

($\mathcal{C}0$) $\emptyset \notin \mathcal{C}$,

($\mathcal{C}1$) $\mathcal{C} = -\mathcal{C}$,

($\mathcal{C}2$) for all $\sigma, \tau \in \mathcal{C}$, if $\operatorname{supp}(\sigma) \subseteq \operatorname{supp}(\tau)$, then $\sigma = \tau$ or $\sigma = -\tau$,

($\mathcal{C}3$) for all $\sigma, \tau \in \mathcal{C}$ and $e \in \underline{\sigma}^+ \cap \underline{\tau}^-$ there exists $\rho \in \mathcal{C}$ such that $\underline{\rho}^+ \subseteq (\underline{\sigma}^+ \cup \underline{\tau}^+) \setminus \{e\}$ and $\underline{\rho}^- \subseteq (\underline{\sigma}^- \cup \underline{\tau}^-) \setminus \{e\}$. (*weak elimination*)

There are other equivalent definitions of matroids and oriented matroids, see [55, 13]. In fact, the possibility to switch between the many different viewpoints is one of the key aspects of matroid theory.

Any oriented matroid $(\mathcal{X}, \mathcal{C})$ defines an ordinary matroid via $C = \{\underline{\sigma} : \sigma \in \mathcal{C}\}$. Conversely, an oriented matroid can be seen as a matroid with an additional structure: This additional structure is called a *circuit orientation* and maps each circuit $\mathcal{Y} \in C$ to a sign vector σ with support $\operatorname{supp}(\sigma) = \mathcal{Y}$.

Let $(\mathcal{X}, C)$ be a matroid. Two elements $x, y \in \mathcal{X}$ are *parallel* if $\{x, y\}$ is a circuit. x and y are *coparallel* if there is no circuit $\mathcal{Y} \in C$ such that $x \in \mathcal{Y}$ and $y \notin \mathcal{Y}$. It is easy to see that these two notions are dual to each other. The elimination axiom (C3)

implies that this defines two equivalence relations that partition $\mathcal{X}$ into *parallel classes* and *coparallel classes*. By definition, a parallel class has rank one. Dually, a coparallel class has corank one.

Two elements $x, y \in \mathcal{X}$ are *connected* if there exists a circuit $\mathcal{Y} \in C$ such that $x, y \in \mathcal{Y}$. This defines a partition of $\mathcal{X}$ into *connected components*.

There are several computer programs that can handle matroids and oriented matroids, at least in the algebraic case, when $\mathcal{N}$ is spanned by $\mathcal{N}_\mathbb{Z} = \mathcal{N} \cap \mathbb{Z}^\mathcal{X}$. For example, `4ti2` [1] can compute a circuit basis. The free software package `TOPCOM` [58] computes the signed circuits and cocircuits of a vector space. Unfortunately, none of these packages can compute all the sign vectors (yet). There are two possibilities to compute the set of all sign vectors from the oriented matroid:

1. Since every sign vector is a composition of signed circuits, the set of all sign vectors can be computed iteratively from the signed circuits by composition.

2. Check every $\sigma \in \{0, \pm 1\}^\mathcal{X}$ whether it satisfies $\sigma \perp \tau$ for all cocircuits τ.

Fortunately, `TOPCOM` provides a library interface to its routines, and the sources (in C++) are freely available under the GPLv2 [29] and clearly written. Moreover, these libraries contain elementary routines for handling symmetry groups. The symmetry groups can be defined by specifying how a set of generators acts on the ground set. This makes it straightforward to implement both algorithms in a way that takes account of the symmetries.

In the general case, when $\mathcal{N}$ is not spanned by $\mathcal{N}_\mathbb{Z}$, there seems to be no computer algebra system that can compute the corresponding (oriented) matroid. One problem is that the oriented matroid is a discrete structure (the number of different oriented matroids with fixed ground set $\mathcal{X}$ is finite), therefore the mapping from linear subspaces $\mathcal{N}$ to oriented matroids cannot be continuous in any sense which would allow for numerical approximations. On the other hand, the ideas behind the algorithms of `4ti2` and `TOPCOM` do not require an integral basis of $\mathcal{N}$. Therefore, it is possible to implement these algorithms in any computer algebra system that can represent a basis of $\mathcal{N}$. There are two reasons why `4ti2` and `TOPCOM` prefer to work over the integers: First, on a computer calculations with integers are much faster, and second, integers are sufficient for most applications.

Bibliography

[1] 4ti2 team, "4ti2—a software package for algebraic, geometric and combinatorial problems on linear spaces," available at www.4ti2.de.

[2] Amari, S., "Information geometry on hierarchy of probability distributions," *IEEE Transactions on Information Theory*, vol. 47, no. 5, pp. 1701–1711, 2001.

[3] Amari, S., and Nagaoka, H., *Methods of Information Geometry*, 1st ed., ser. Translations of mathematical Monographs. American Mathematical Society, 2000, vol. 191.

[4] Ay, N., "Information geometry on complexity and stochastic interaction," Max Planck Institute for Mathematics in the Sciences, Leipzig, Preprint 95, 2001.

[5] ——, "An information-geometric approach to a theory of pragmatic structuring," *Annals of Probability*, vol. 30, pp. 416–436, 2002.

[6] ——, "Locality of global stochastic interaction in directed acyclic networks," *Neural Computation*, vol. 14, pp. 2959–2980, 2002.

[7] Ay, N., and Knauf, A., "Maximizing multi-information," *Kybernetika*, vol. 42, pp. 517–538, 2006.

[8] Ay, N., Olbrich, E., Bertschinger, N., and Jost, J., "A unifying framework for complexity measures on finite systems," in *Proceedings ECCS'06*, 2006, Santa Fe Institute Working Paper 06-08-028.

[9] Ay, N., and Wennekers, T., "Dynamical properties of strongly interacting Markov chains," *Neural Networks*, vol. 16, no. 10, pp. 1483–1497, 2003.

[10] Barndorff-Nielsen, O., *Information and Exponential Families in Statistical Theory*, 1st ed. Wiley, 1978.

[11] Bates, D., Hauenstein, J., Sommese, A., and Wampler, C., "Bertini: Software for numerical algebraic geometry," available at http://www.nd.edu/~sommese/bertini.

[12] Bi, G.-Q., and Poo, M.-M., "Synaptic modification by correlated activity: Hebb's postulate revisited," *Annual Review of Neuroscience*, vol. 24, no. 1, pp. 139–166, 2001.

[13] Björner, A., Las Vergnas, M., Sturmfels, B., White, N., and Ziegler, G., *Oriented Matroids*, 1st ed., ser. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1993.

[14] Bogart, T., Jensen, A., and Thomas, R., "The circuit ideal of a vector configuration," *Journal of Algebra*, vol. 309, no. 2, pp. 518–542, 2007.

[15] Brown, L., *Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory*. Hayworth, CA, USA: Institute of Mathematical Statistics, 1986.

[16] Chentsov, N. N., *Statistical Decision Rules and Optimal Inference*, 1st ed., ser. Translations of Mathematical Monographs. American Mathematical Society, 1982, Russian original: Nauka, Moscow, 1972.

[17] Cover, T., and Thomas, J., *Elements of Information Theory*, 1st ed. Wiley, 1991.

[18] Cox, D., Little, J., and Schenck, H., *Toric Varieties*, 1st ed. AMS, 2011, preliminary version available at http://www3.amherst.edu/~dacox/ until publication.

[19] Cox, D. A., Little, J., and O'Shea, D., *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*, 3rd ed. Springer, 2008.

[20] Csiszár, I., and Shields, P., *Information Theory and Statistics: A Tutorial*, 1st ed., ser. Foundations and Trends in Communications and Information Theory. now Publishers, 2004.

[21] Csiszár, I., and Matúš, F., "Closures of exponential families," *Annals of Probability*, vol. 33, pp. 582–600, 2005.

[22] ——, "Generalized maximum likelihood extimates for exponential families," *Probability Theory and Related Fields*, vol. 141, pp. 213–246, 2008.

[23] Darroch, J., and Ratcliff, D., "Generalized iterative scaling for log-linear models," *The Annals of Mathematical Statistics*, vol. 43, no. 5, pp. 1470–1480, 1972.

[24] Della Pietra, S., Della Pietra, V., and Lafferty, J., "Inducing features of random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 380–393, 1997.

[25] Diaconis, P., and Sturmfels, B., "Algebraic algorithms for sampling from conditional distributions," *Annals of Statistics*, vol. 26, pp. 363–397, 1998.

[26] Drton, M., Sturmfels, B., and Sullivant, S., *Lectures on Algebraic Statistics*, 1st ed., ser. Oberwolfach Seminars. Birkhäuser, Basel, 2009, vol. 39.

[27] Eisenbud, D., and Sturmfels, B., "Binomial ideals," *Duke Mathematical Journal*, vol. 84, no. 1, pp. 1–45, 1996.

[28] Endrass, S., "surf 1.0.4," available at http://surf.sourceforge.net, 2003.

[29] Free Software Foundation, "GNU General Public License version 2," available at http://www.gnu.org/licenses/old-licenses/gpl-2.0.txt, 1991.

[30] Fulton, W., *Introduction to Toric Varieties*, 1st ed. Princeton University Press, 1993.

[31] Geiger, D., Meek, C., and Sturmfels, B., "On the toric algebra of graphical models," *Annals of Statistics*, vol. 34, no. 5, pp. 1463–1492, Oct 2006.

[32] Grayson, D., and Stillman, M., "Macaulay2, a software system for research in algebraic geometry," available at http://www.math.uiuc.edu/Macaulay2/.

[33] Greuel, G.-M., and Pfister, G., *A Singular Introduction to Commutative Algebra*, 1st ed. Springer, 2002.

[34] Greuel, G.-M., Pfister, G., and Schönemann, H., "Singular 3-1-2 — A computer algebra system for polynomial computations," Available at http://www.singular.uni-kl.de, 2010.

[35] Grünbaum, B., *Convex Polytopes*, 2nd ed., ser. Graduate Texts in Mathematics. Springer, 2003, second edition prepared by V. Kaibel, V. Klee and G. Ziegler.

[36] Hemmecke, R., and Malkin, P., "Computing generating sets of lattice ideals and Markov bases of lattices," *Journal of Symbolic Computation*, vol. 44, pp. 1463–1476, 2009.

[37] Hertz, J., Krogh, A., and Palmer, R., *Introduction to the Theory of Neural Computation*, ser. Santa Fe Institute Studies In The Sciences Of Complexity Lecture Notes. Addison-Wesley, 1991, vol. 1.

[38] Hoşten, S., and Sullivant, S., "Gröbner bases and polyhedral geometry of reducible and cyclic models," *Journal of Combinatorial Theory: Series A*, vol. 100, no. 2, pp. 277–301, april 2002.

[39] Hodgkin, A., and Huxley, A., "A quantitative description of membrane current and its application to conduction and excitation in nerve," *Bulletin of Mathematical Biology*, vol. 52, no. 1, pp. 25–71, 1952.

[40] Jaynes, E. T., "Information theory and statistical mechanics," *The Physical Review*, vol. 106, no. 4, pp. 620–630, 1957.

[41] Juríček, J., "Maximization of information divergence from multinomial distributions," *Acta Universitatis Carolinae*, vol. 52, no. 1, 2011, in press.

[42] Kahle, T., Olbrich, E., Jost, J., and Ay, N., "Complexity measures from interaction structures," *Physical Review E*, vol. 79, no. 2, p. 026201, 2009.

[43] Kahle, T., "Decomposition of binomial ideals," *Annals of the Institute of Statistical Mathematics*, vol. 62, no. 4, pp. 727–745, 2010.

[44] Katsabekis, A., and Thoma, A., "Parametrizations of toric varieties over any field," *Journal of Algebra*, no. 308, pp. 751–763, 2007.

[45] Kimball, S., Mattis, P., and others, "gimp 2.6," available at http://www.gimp.org, 2010.

[46] Lauritzen, S. L., *Graphical Models*, 1st ed., ser. Oxford Statistical Science Series. Oxford University Press, 1996.

[47] Linsker, R., "Self-organization in a perceptual network," *IEEE Computer*, vol. 21, pp. 105–117, 1988.

[48] Malkin, P., "Computing Markov bases, Gröbner bases, and extreme rays," Ph.D. dissertation, Université de Louvain, 2007.

[49] Matúš, F., "Maximization of information divergences from binary i.i.d. sequences," *Proceedings of IPMU*, vol. 2, pp. 1303–1306, 2004.

[50] ——, "Optimality conditions for maximizers of the information divergence from an exponential family," *Kybernetika*, vol. 43, no. 5, pp. 731–746, 2007.

[51] ——, "Divergence from factorizable distributions and matroid representations by partitions," *IEEE Transactions in Information Theory*, vol. 55, pp. 5375–5381, 2009.

[52] Matúš, F., and Ay, N., "On maximization of the information divergence from an exponential family," in *Proceedings of the WUPES'03*. University of Economics, Prague, 2003, pp. 199–204.

[53] Matúš, F., and Rauh, J., "Maximization of the information divergence from an exponential family and criticality," in *2011 IEEE International Symposium on Information Theory Proceedings (ISIT2011)*, St. Petersburg, Russia, Jul. 2011.

[54] McCulloch, W., and Pitts, W., "A logical calculus of the ideas immanent in nervous activity," *Bulletin of Mathematical Biophysics*, vol. 5, pp. 115–133, 1943.

[55] Oxley, J., *Matroid Theory*, 1st ed. New York: Oxford University Press, 1992.

[56] Pfister, G., Decker, W., Schönemann, H., and Laplagne, S., "`primdec.lib`. A singular 3-1-2 library for computing the normalization of affine rings," 2010.

[57] Pistone, G., Riccomagno, E., and Wynn, H., *Algebraic Statistics: Computational Commutative Algebra in Statistics*, ser. Monographs on Statistics and Applied Probability. Chapman and Hall, CRC Press, 2001.

[58] Rambau, J., "TOPCOM: Triangulations Of Point Configurations and Oriented Matroids," in *Mathematical Software—ICMS 2002*, Cohen, A., Gao, X.-S., and Takayama, N., Eds. World Scientific, 2002, pp. 330–340.

[59] Rao, C. R., *Linear Statistical Inference and its Applications*, 2nd ed., ser. Wiley Series in Probability and Statistics. Wiley, 1973.

[60] Rauh, J., "Finding the maximizers of the information divergence from an exponential family," *IEEE Transactions on Information Theory*, vol. 57, no. 6, pp. 3236–3247, 2011.

[61] Rauh, J., Kahle, T., and Ay, N., "Support sets of exponential families and oriented matroids," *International Journal of Approximate Reasoning*, vol. 52, no. 5, pp. 613–626, 2011.

[62] Schuster, P., "How does complexity arise in evolution," *Complexity*, vol. 2, pp. 22–30, September 1996.

[63] Shannon, C., "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423 and 623–656, 1948.

[64] Smith, J. G., "The information capacity of amplitude- and variance-constrained scalar gaussian channels," *Information and Control*, vol. 18, no. 3, pp. 203–219, 1971.

[65] Sommese, A., and Wampler, C., *The Numerical Solution of Systems of Polynomials Arising in Engineering and Science*, 1st ed. World Scientific Publishing Company, 2005.

[66] Sporns, O., *Networks of the Brain*, 1st ed. The MIT Press, 2011.

[67] Steiner, L., and Kahle, T., "cipi—computing information projections iterativly," available at http://personal-homepages.mis.mpg.de/kahle/cipi.

[68] Studený, M., "Asymptotic behaviour of empirical multiinformation," *Kybernetika*, vol. 23, no. 2, pp. 124–135, 1987.

[69] Sturmfels, B., *Gröbner Bases and Convex Polytopes*, 1st ed. American Mathematical Society, 1996.

[70] Tononi, G., Sporns, O., and Edelman, M., "A measure for brain complexity: Relating functional segregation and integration in the nervous systems," *Proceedings of the National Acadademy of Science USA*, vol. 91, pp. 5033–5037, 1994.

[71] Weis, S., "Exponential families with incompatible statistics and their entropy distance," Ph.D. dissertation, Friedrich-Alexander-Universität Erlangen-Nürnberg, 2009.

[72] Wennekers, T., and Ay, N., "Finite state automata resulting from temporal information maximization and a temporal learning rule," *Neural Computation*, vol. 17, no. 10, pp. 2258–2290, 2005.

[73] Williams, T., Kelley, C., and others, "gnuplot 4.4.2," available at http://www.gnuplot.info, 2010.

[74] Zhu, S. C., Wu, Y. N., and Mumford, D., "Minimax entropy principle and its application to texture modeling," *Neural Computation*, vol. 9, pp. 1627–1660, November 1997.

[75] Ziegler, G., *Lectures on Polytopes*, 2nd ed. Springer, 1998.

[76] Zvárová, J., "On measures of statistical dependence," *Časopis pro pěstování matematiky*, vol. 99, 1974.

Index

Page numbers *in italics* refer to definitions.

Glossary of notations

This glossary contains a list of symbols that are used throughout this thesis. Numbers refer to the pages that containing a definition of the symbol.

$\lceil a \rceil$	The ceiling function	5
$\lfloor a \rfloor$	The floor function	5
$[n]$	A set of cardinality n	
$\mathbf{1}$	The uniform measure	
$\mathbf{1}_{\mathcal{Y}}$	The characteristic function on $\mathcal{Y}$	
A	The sufficient statistics matrix	9
a_i	A row of the sufficient statistics matrix	
A_x	A column of the sufficient statistics matrix	
$\mathrm{Bin}(n)$	The binomial model	82
$\overline{D}_{\mathcal{E}}$	The function $\overline{D}_{\mathcal{E}}$	35
$\Delta_{P,Q}$	The line segment between P and Q	
$\Delta^{\circ}_{P,Q}$	The (relative) interior of the line segment between P and Q	
δ_x	A point measure concentrated at x	
$D(P\|\mathcal{E})$	The information divergence from P to $\mathcal{E}$	11
$\mathcal{E}$	An exponential family	7
$\mathcal{E}_{(2)}$	The pure paire interaction exponential family	23
$\mathcal{E}_{\Delta}$	The hierarchical exponential family of Δ	21
$\overline{\mathcal{E}^n_{\mathrm{iid}}}$	The binary i.i.d. model	82
$\overline{\mathcal{E}_{\nu,\mathcal{T}}}$	The closure of the exponential family with reference measure ν and tangent space $\mathcal{T}$	
$\mathcal{H}_{\mathbf{1}}$	The set of exponential families with uniform reference measure	
$\mathfrak{i}$	The imaginary unit	
$\mathrm{K}_{\mathcal{E}}$	The set of kernel distributions of $\mathcal{E}$	30
$\mathbf{M}_A$	The convex support	9
$\mathcal{N}$	The normal space	8

$\mathcal{N}_P$	The linear family of P	11
ν	The reference measure	7
$P_{\mathcal{E}}$	The rI-projection of P onto $\mathcal{E}$	12
π_A	The moment map	9
π_Δ	The Δ-margins	22
π_S	The S-margins	22
$\Psi_{\mathcal{E}}$	A map $\mathbf{P}(\mathcal{X}) \setminus \mathcal{E} \to \partial \mathbf{U}_{\mathcal{N}}$	28
Ψ^+	The map $u \mapsto u^+$	35
$\mathbf{P}(\mathcal{X})$	The set of probability measures on $\mathcal{X}$	6
$\mathbf{P}(\mathcal{X})^\circ$	The set of strictly positive probability measures on $\mathcal{X}$	6
$\mathbb{R}_\geq$	The nonnegative reals	
$\mathcal{T}$	The tangent space	7
$\tilde{\mathcal{T}}$	The extended tangent space	8
$\mathbf{U}_{\mathcal{N}}$	The polytope of differences of probability measures in $\mathcal{N}$	31
$\mathcal{X}$	A finite set, the ground set	

List of Figures

Note: All figures were created with gnuplot [73], with one exception: Figure 3.4 was created with surf [28] and post-processed using gimp [45].

Acknowledgements

I thank Jürgen Jost for giving me the opportunity for my doctoral studies at the MPI MIS at Leipzig. His support gave me the freedom to do what I wanted to do and to choose my own direction of research.

I thank Nihat Ay for proposing the main subject of my thesis. He always had time when I needed his advice and feedback. I thank him for his trust and his support of my plans.

I am grateful to Thomas Kahle, who infected me with his enthusiasm for the interplay of combinatorics, information theory and algebra. I value the many conversations with him about math, the universe and everything and his remarks about my manuscripts.

I thank Fero Matúš for many fruitful discussions. I learnt a lot during my many visits to Prague. His pursuit of utmost logical clarity helped me to structure my ideas and to deepen my understanding.

I am grateful to Bernd Sturmfels, whose email got me started on the subject. He taught me a lot and inspired me to learn even more.

I thank Bastian Steudel for several stimulating discussions and for helping me with computer simulations.

I thank Antje Vandenberg for her assistance and support in all organizational aspects and for solving many problems even before I discovered them. I also thank the computer group and the administrative staff for their quick response to any issues.

I thank the IMPRS and the RAL for their generous funding of my various travels, most notably to the Czech Republic.

I also thank all other people at the institute for creating this inspiring atmosphere, with many fruitful discussions, and for ensuring the continuous supply of tea and cookies.

I thank my parents for patiently supporting my studies.

Finally, I thank Elisabeth for her enduring patience and continuous support.

About the author

Name:	Johannes Rauh
Date of birth:	30th September 1981 in Würzburg
2000–2007	Studies in Physics and Mathematics at the University of Würzburg
2006	Diplom in Physics
2007	Diplom in Mathematics
2007–2011	Doctoral studies in the IMPRS at the Max Planck Institute for Mathematics in the Sciences at Leipzig

Bibiliographische Daten

Finding the Maximizers of the Information Divergence from an Exponential Family
(Das Auffinden der Maximierer der Informationsdivergenz
von einer Exponentialfamilie)
Rauh, Johannes
Universität Leipzig, Dissertation, 2011
125 Seiten, 12 Abbildungen, 76 Referenzen, eine Fußnote.

Changes in the final version

This section summarizes the changes that have been made from the version that was submitted in March 2011 to the Mathematical Institute of the University of Leipzig. Minor changes, such as typos, are omitted.

Some of the results of Chapter 3 have meanwhile been published; the corresponding publications are now referenced.

Theorem 5.5 has been slightly generalized and now also applies to exponential families that do not contain the uniform distribution. This makes most other results in Section 5.2 that follow Theorem 5.5 more general.