On the Expressive Power of Discrete Mixture Models, Restricted Boltzmann Machines, and Deep Belief Networks—A Unified Mathematical Treatment

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. Guido Francisco Montúfar Cuartas,
geboren am 13.05.1983 in Panama-Stadt (Panama).

Die Annahme der Dissertation haben empfohlen:

1. Professor Dr. Jürgen Jost (MPI MIS Leipzig)
2. Professor Dr. Christoph Schnörr (Universität Heidelberg)

Die Verleihung des akademischen Grades erfolgt mit Bestehen der
Verteidigung am 17.10.2012
mit dem Gesamtprädikat magna cum laude.

Contents

List of Figures

Figures 1.2, 1.3, 3.1, 3.2, 4.1, 4.4, and 5.4 were created using PSTricks. All other figures were created in MATLAB and annotated in LaTeX.

Introduction

Motivation and Previous Results

This thesis is about the representational power of discrete statistical models underlying artificial learning machines. The main contribution of this work is to the estimation of the number of variables needed to obtain sufficiently rich models and learning systems that comply prescribed learning capabilities. An emphasis is placed on the models of type Restricted Boltzmann Machine and Deep Belief Network, which have played a key role in the quest for Artificial Intelligence (AI) in the last few years. The representational power of statistical models is directly related to goodness of fit and model complexity in model selection, and to geometric and combinatorial properties; such as the dimension, boundary, convex hull, and approximation errors of the set of probability distributions that they comprise. On that account, this work follows a geometric-combinatorial approach. The geometric-combinatorial approach to statistical models and learning theory is notably represented by the theory of *information geometry* [117, 4, 8, 10] developed by S. Amari and coworkers, and the emerging field *algebraic statistics* [41, 48]. It has been applied successfully in experimental design, statistical inference, hypothesis testing, neural networks, cognitive systems, statistical physics and computational biology.

The central goal of the machine learning research field *Deep Learning* is to obtain abstract representations of data that can be used for AI-related tasks; including data analysis, classification, pattern recognition, perception, machine vision and speech recognition. Deep neural networks emulate cognitive processes in the human brain through several levels of non-linear information processing and hierarchical feature extraction [115, 19]. Training deep architectures is difficult; for example, exact maximum likelihood (ML) gradient methods for classical Boltzmann Machines [2, 59, 9] require exceedingly expensive evaluation of expectation functions. Several alternative learning algorithms and deep architectures have been proposed to overcome these problems. In 2006 G. E. Hinton, S. Osindero and Y. Teh [58] proposed an efficient greedy, layer-wise, unsupervised training algorithm for a new architecture: the *Deep Belief Network* (DBN). Their method represented a breakthrough that boosted many results and successful applications in a number of tasks; e.g., dimensionality reduction, information retrieval, speech recognition, modeling, machine vision, classification. See Y. Bengio [17] for a complete overview. The DBN has a graphical representation with visible nodes at the bottom of a network with several layers of binary units[1], where only units in subsequent layers are connected and all connections are directed towards the visible layer, except for the connections between the deepest two layers, which are undirected. The DBN is inherently related to the *Restricted Boltzmann Machine* (RBM) (P. Smolensky [107], Y. Freund and D. Haussler [45], G. E. Hinton [56]). The RBM is a learning system with one hidden and one visible layer of binary stochastic units, and a complete bipartite undirected graph of interactions between them. The RBM has an efficient unsupervised training algorithm called *Contrastive Divergence* (CD) [56, 25]. The RBM and its training algorithms are used to progressively train the parameters of a DBN [58, 18]. Train-

[1] Many extensions and generalizations of DBNs have been proposed, including models with continuous variables. We focus on the standard, most important, binary DBN.

ing is based on the conditional independence assumptions (all visible units are conditionally independent given the state of the hidden units and vice versa) represented by the connectivity constraints (see [84] for comments on this).

A fundamental question is whether deep learning systems (i.e., systems involving several layers of variables) are better than shallow systems at solving AI-related tasks. This motivates the following question: *What classes of marginal probability distributions can be represented by deep and shallow networks, respectively?* For RBMs and DBNs the various notions of model dimension, complexity, and errors are far from being fully understood. The sets of probability distributions that can be represented by these models build intricate geometric objects depending on the number of hidden units and layers that they contain. Y. Freund and D. Haussler [45], and N. Le Roux and Y. Bengio [72] showed that RBMs are universal approximators, provided they have enough hidden units. Recently M. Aoyagi [11] computed bounds for the asymptotic generalization error of RBMs (disregarding some of the model parameters and assuming the *true* probability distributions are contained in the model), within S. Watanabe's framework of *singular learning theory* [119]. A. Cueto, J. Morton and B. Sturmfels [31] used *tropical geometry*, a newly developed method of algebraic geometry, to show that the RBM model has the *expected dimension*, (i.e., it has dimension equal to the number of model parameters or to the dimension of the ambient probability simplex, whichever is smaller), if the number of hidden units is small enough or large enough. One of the most essential questions on DBN research has been the following: *Does a DBN exist which can approximate any distribution on the states of the visible units through appropriate choice of parameters?* Furthermore: *If such DBN* universal approximators *exist, what is their minimal size? What is the tradeoff between the size of the hidden layers and the number of hidden layers?* I. Sutskever and G. E. Hinton [110] showed that a very deep and narrow DBN, consisting of $\sim 3 \cdot 2^n$ hidden layers of width $(n+1)$, can approximate any distribution on $\{0,1\}^n$ arbitrarily well. N. Le Roux and Y. Bengio [73] improved this bound showing that $\sim \frac{2^n}{n}$ layers of width n suffice. In Chapter 3 we improve all bounds for RBMs and DBNs and thereby resolve a conjecture that was posed in [73].

Given the restricted connectivities of DBNs and RBMs, universal representation of probability distributions necessarily requires a number of hidden units which is exponential in the number of visible units, in correspondence to the dimension of the visible probability simplex ($2^n - 1$ for n visible binary units). In applications it is desired to represent only some parts of the probability simplex. In particular, real world systems often confine to a relatively small set of their state space (think of natural images or written English). Hence it is important to understand the representational power of learning systems in terms of selected classes of probability distributions. In Chapters 4 and 5 we study submodels of RBMs and DBNs and present suitable classes of probability distributions that can be learned by these systems depending on the number of hidden variables that they contain. Furthermore, we bound the model approximation errors when approximating arbitrary target probability distributions and target probability distributions from selected classes.

Models with latent variables describe probability distributions of the following form: $p(v|\theta) = \sum_h p(v|h;\theta)p(h|\theta)$, where h is the state of the hidden variables and θ is the model parameter. They can be understood as mixture distributions with mixture weights $p(h|\theta)$ and mixture components $p(\cdot|h;\theta)$ determined by the model parameter θ and restricted through the model assumptions. In a *mixture model* the only restriction is that h belongs to a fixed set and all $p(\cdot|h;\theta)$

belong to a common model. We exploit the relationships between latent variable models and mixture models. The problem of representing probability distributions as mixtures with specific properties has a long history. A prominent example is a result by de Finetti which states that infinite exchangeable distributions are mixtures of independent and identically distributed Bernoulli sequences[2]. Mixture models find numerous applications; e.g., modeling of heterogeneous populations, clustering and machine learning. There is an abundant literature on the subject, for instance [75, 80, 114, 20]. The *expectation-maximization* (EM) algorithms [33] provide tools for density estimation. There is a diversity of interesting results on mixture models; e.g., dispersion of mixtures [104], the method of moments [92], parameter identifiability [22].

The expressive power of mixture models is not sufficiently well understood and continues to be an extremely active field of research. *How many mixture components from a discrete* exponential family *(log-linear model) are required to represent or to approximate distributions from a more complicated family?* How many latent states are required in order to explain a stochastic experiment? This problem can be formulated as follows: Let $\mathrm{Mixt}^m(\mathcal{M})$ denote the set of convex combinations of any m elements from the set $\mathcal{M}$ and let $\mathrm{conv}(\mathcal{M})$ denote the convex hull of $\mathcal{M}$. *Given two exponential families $\mathcal{E}$ and $\mathcal{E}'$ with a finite state space $\mathcal{X}$ and $\mathcal{E}' \subseteq \mathrm{conv}(\mathcal{E})$, find the natural number $f = f(\mathcal{E}, \mathcal{E}')$ that satisfies the relation $\mathrm{Mixt}^m(\mathcal{E}) \supseteq \mathcal{E}'$ if and only if $m \geq f$.* In convex analysis the number f is referred to as the *Carathéodory number* of $\mathcal{E}'$ with respect to $\mathcal{E}$. The case where $\mathcal{E}'$ equals the entire set of probability distributions $\mathcal{P}(\mathcal{X})$ corresponds to finding the smallest m (if there is any) for which the mixture model is a *universal approximator*. An associated problem is to find the Carathéodory number of $\mathcal{E}$, i.e., the smallest m for which $\mathrm{Mixt}^m(\mathcal{E}) = \mathrm{conv}(\mathcal{E})$. If $\mathcal{E}$ is the set of product distributions of n variables, then $f(\mathcal{E}, \mathcal{E}')$ is the maximum non-negative outer-product rank of the n-way tables of probabilities (tensors) described by $\mathcal{E}'$. For the set of product distributions with two variables $\mathcal{E}^1(\mathcal{X}_1 \times \mathcal{X}_2)$, it is well known that $m \geq \min\{|\mathcal{X}_1|, |\mathcal{X}_2|\}$ implies $\mathrm{Mixt}^m(\mathcal{E}^1(\mathcal{X}_1 \times \mathcal{X}_2)) = \mathcal{P}$, see [103, 7]. This observation is equivalent to stating that every non-negative $k \times l$ matrix can be written as the non-negative sum of at most $\min\{k, l\}$ rank-one matrices. It is also known that if $|\mathcal{X}_1|, |\mathcal{X}_2| > 2$, then $\mathrm{Mixt}^2(\mathcal{E}^1(\mathcal{X}_1 \times \mathcal{X}_2)) \neq \mathcal{P}$, see [50]. Hence $\min\{|\mathcal{X}_1|, |\mathcal{X}_2|\} \geq f(\mathcal{E}^1(\mathcal{X}_1 \times \mathcal{X}_2), \mathcal{P}) > 2$ when $|\mathcal{X}_1|, |\mathcal{X}_2| > 2$. In Chapter 1 we generalize these results; for instance, we provide exact values for $f(\mathcal{E}^1(\mathcal{X} \times \cdots \times \mathcal{X}), \mathcal{P})$.

The boundary of an exponential family is a union of exponential families supported by characteristic subsets of its state space. Computing the support sets is a difficult combinatorial problem equivalent to describing the face lattice of the convex support polytope [16] and has interesting relations to oriented matroids and coding theory, see [64, 96]. If an entire face of the probability simplex is contained in a statistical model, we call the support of that face an S-set of the model (S for simplex). The S-sets of exponential families are simplex faces of their convex supports. Hence the Carathéodory number is bounded from above by the cardinality of a covering of the vertices of the convex support polytope by simplex faces. In Chapter 1 we derive results for this kind of coverings for marginal polytopes of hierarchical models. An important idea in this thesis is to investigate models arising from mixing distributions supported by extreme points (sets) of exponential families, in analogy to *Choquet theory*. The analysis of extreme points is a powerful tool to the study convex geometric objects (think of ergodic decompositions of stationary processes, pure states in quantum mechanics, deterministic policies, or just Carathéodory's

[2]P. Diaconis [36] shows approximation results in the case of finite exchangeable sequences.

theorem). In Chapter 2 we investigate a variety of related topics, and Chapters 3 and 4 deal with mixtures of product distributions with disjoint supports, which are mixtures based at the boundary of an independence model.

Main Results and Outline

In the following we give an informal description of the main results of this thesis.

Part I. Exponential Families and Mixture Models

Chapter 1 (Mixtures of Discrete Exponential Families). A combinatorial approach to the representational power of mixtures of discrete multivariate exponential families is proposed based on combinatorics of convex polytopes and coding theory. We introduce a special class of support sets of statistical models, the S-sets, and relate the Carathéodory numbers to coverings and packings of support sets.

The problem of finding the minimal number of mixture components from an exponential family needed to represent any probability distribution is intrinsically related to the geometry of the exponential family. In particular, the dimension of the set of mixtures must equal the dimension of the ambient probability simplex. The most important special case is the mixture of product distributions. Until recently it was a long standing problem whether the set of mixtures of m products of n binary distributions had the *expected dimension* $\min\{n \cdot m + (m-1), 2^n - 1\}$ for all m and n. M. Catalisano, A. Geramita, and A. Gimigliano [26] proved that this model has the expected dimension, unless $n = 4$ and $m = 3$. For mixtures of non-binary product distributions the problem is still open.

In the main results of this chapter we find bounds on the Carathéodory numbers of hierarchical models expressed in terms of the number of variables and the cardinality of their state spaces. For product distributions we establish the following:

The smallest m for which any probability distribution on $\{1, \ldots, q\}^n$ can be represented as the mixture of m product distributions is q^{n-1} (where q is any prime power).

The specified number q^{n-1} is larger than expected; in the binary case the mixture model has the same dimension as the probability simplex when $m \geq \frac{2^n}{(n+1)}$, but universal approximation requires $m \geq 2^{n-1}$ and at least $(n+1)2^{n-1} - 1$ parameters. Mixtures of binary product distributions are important in our analysis of RBMs and DBNs in Chapters 3 and 4.

For probability distributions involving interactions among any k variables we show that:

The smallest m for which any probability distribution on $\{0, 1\}^n$ can be represented as the mixture of m distributions from the k-interaction model is not more than $2^{n-(k+1)}(1 + \frac{1}{(2^k-1)})$.

This result is derived from coverings of the vertex set of the convex support of k-interaction models by simplex faces. A full characterization of the simplex faces and the computation of the smallest packing for general k and for non-binary variables remains a challenging problem. The mixture models of k-interaction models have a hierarchical representation with a latent variable. This allows us to define a stochastic dynamics on the states of its variables. By our result, it is possible to control the expressive power of such generalized stochastic networks which include

higher-order interactions.

In addition to the results outlined above, we derive bounds for the minimal number of mixtures of binary product distributions needed to represent k-interaction models. Furthermore, we study the number of modes of mixtures of binary product distributions (in the graph of the n-hypercube), derive inequality constraints for these models, and bound the volume of their complement in the probability simplex from below.

Chapter 2 (Convex Subsets, Secants, Geodesics and Convex Hulls). We explore geometric properties of exponential families motivated by the following problem: *Find the smallest m for which* $\mathrm{Mixt}^m(\mathcal{E}) = \mathrm{conv}(\mathcal{E})$. This problem gives rise to numerous related questions, including the following: *How much do we learn about mixtures of exponential families when we study only the mixtures with basis points at the boundary of the exponential families? What is the maximal dimension of a convex subset of an exponential family?* Many interesting topics arose in the process of developing tools to solve the just motivated questions. This chapter contains a variety of individual results. The implications of some results to the main subject of this thesis are not fully elaborated at this moment and should be understood as a basis for future research. In Section 2.1 we elaborate on convex subsets of exponential families and relate them to the convex supports. Furthermore, we characterize convex α-families, extending previous results by F. Matúš and N. Ay for exponential families. Section 2.2 studies secant lines of exponential families and shows that:

The intersections of lines and exponential families encode portions of their support set lattices and their convex subfamilies.

These results imply, in particular, that an intersection index encodes the existence of convex decompositions of exponential families (more precisely, foliations into simplices of a certain dimension), and that an exponential family which intersects a generic line at $(d+1)$ points contains every $\lfloor \frac{d}{2} \rfloor$-dimensional face of the probability simplex (or equals the full probability simplex). For example, any binary independence model intersects a line at either $0, 1, 2$ or ∞ points and is an n-ruled manifold. Furthermore, the 2-bit independence model is not a minimal surface.

In Section 2.4 we use these new methods to compute the Carathéodory number of some classes of non-hierarchical exponential families, derive sufficient and necessary conditions for exponential families to be contained in the convex hull of their boundaries, and analyze the convex hulls of α-geodesics. We investigate α-mixtures of exponential families targeting an interpolation between exponential families and their mixture models.

Part II. Restricted Boltzmann Machines and Deep Belief Networks

Chapter 3 (Universal Approximation Results for RBMs and DBNs). This chapter studies the problem of universal approximation of probability distributions by neural networks of type Restricted Boltzmann Machine and Deep Belief Network. We show that:

An RBM with $\frac{2^n}{2} - 1$ hidden units is capable of approximating any distribution on $\{0,1\}^n$ arbitrarily well as its marginal visible distribution.

This RBM has $(n+1)2^{n-1}-1$ parameters. This result improves the previously known upper bounds for the minimal size of an RBM universal approximator [72, 73]. Furthermore, we show:

A DBN with $\frac{2^n}{2(n-\log_2(n))}$ *hidden layers of size* n *is capable of approximating any distribution on* $\{0,1\}^n$ *arbitrarily well as its marginal visible distribution.*

Such a DBN contains roughly $(n+1)2^{n-1}$ parameters. This result improves the previously known upper bounds for the minimal size of a narrow DBN universal approximator [110, 72, 73].

At this moment we don't know if there exist RBM or DBN universal approximators with fewer parameters. It is striking that our upper bounds for the minimal number of parameters in RBM and narrow DBN universal approximators coincide, and furthermore, that these bounds coincide with the exact minimal number of parameters of a universal approximating mixture of binary product distributions (computed in Chapter 1).

In addition to the above mentioned results, we show that an RBM with three visible and two hidden units is not a universal approximator, although the model has the same dimension as the ambient probability simplex. In turn, the smallest RBM universal approximator with three visible units has three hidden units, corresponding to our bound. For completeness we give a formal discussion of parameter counting bounds for the minimal size of universal approximators. In Appendix 3.C we compare the parametrizations of the probability simplex which arise from RBM and DBN universal approximators.

Chapter 4 (Expressive Power and Approximation Errors of RBMs and DBNs). This chapter presents a hierarchy of explicit classes of probability distributions that RBMs and DBNs can represent expressed in terms of the number of hidden units and hidden layers of the RBMs and DBNs. These classes include large collections of mixtures of product distributions with disjoint supports. If the number of hidden units and layers is large enough, these submodels fill the entire probability simplex in accordance with the results from Chapter 3. The geometry of these submodels is easier to study, while they still capture important properties of the models. Using these results we are able—for the first time—to bound the maximal approximation errors of RBMs and DBNs.

The maximal Kullback-Leibler (KL) divergence from any points in the probability simplex to a model $\mathcal{M}$ is a measure of the model errors. The problem of maximizing the KL-divergence to an exponential family was originally proposed by N. Ay [12] and has been treated by N. Ay, F. Matúš and J. Rauh. In Section 4.1 we show upper and lower bounds for the maximal KL-divergence in the case of *mixtures* of product distributions. In the case of RBMs we show:

It is always possible to reduce the error of an RBM with n *visible and* m *hidden units to at most* $(n-1)-\log(m+1)+0.1$.

Computer experiments showed that the bound captures the order of magnitude of the true approximation error, at least for small RBMs. In particular, for the RBM with three visible and two hidden units the KL-distances from any three-dimensional face of the probability simplex to the model and to the proposed submodel agree, and our bound is exact. We show that the dimension of the RBM model strictly increases with the number of hidden units until reaching the dimension of the ambient simplex, which slightly extends results from [31]. In the case of

DBNs we show:

It is always possible to reduce the error of a DBN with n visible units and l hidden layers of width n to at most $n - \log(2l\log(l))$.

Our approach can be generalized to treat DBNs with layers of different widths. Our results give a theoretical basis for selecting the size of an RBM and a DBN which accounts for a desired model error tolerance. On the other hand, learning may not always find the best approximation, resulting in an error that may well exceed our bound. We believe that our bound for the approximation errors of DBNs can be improved through a more detailed analysis of the proposed submodels.

In addition to the above results, we study the maximal number of modes of distributions from RBM models and show that the smallest mixture model of product distributions which contains an RBM has a much larger dimension:

In essentially all cases of practical interest, an exponentially larger mixture model, requiring an exponentially larger number of parameters, is required to represent the distributions that can be represented by the RBM.

We examine two particular RBM models in detail. We find that the RBM with three visible and two hidden units, however full dimensional, can't represent any distribution with four strong local maxima (locality with respect to the Hamming distance on $\{0,1\}^n$) and provide evidence showing that the uniform distribution is not an inner point of that model.

Chapter 5 (Model Design). The universal approximation problem, as treated in Chapter 3, is to reduce the maximal KL-divergence from arbitrary distributions in the probability simplex to a model $\mathcal{M}$, $\max\{D(p\|\mathcal{M}) \colon p \in \mathcal{P}\}$, to zero as a function of the hyperparameters of the model (the number of hidden units and layers for the classes RBM and DBN). The problem can be treated for specific classes of target distributions $p \in \mathcal{G} \subseteq \mathcal{P}$. The resulting problem is called *Model Design*. In many cases there are no explicit descriptions of the interesting classes of targets. In fact, one of the motivations for training DBNs is to obtain simpler representations of the target distributions. Often however, certain properties of the targets are known in advance.

In Section 5.1 we study approximation errors when approximating distributions from particular classes of interest. For example, we consider the problem of representing deterministic kernels by RBMs. Section 5.2 discusses how to reduce the search space of learning systems. We use exponential families to parametrize objects other than probability distributions in such a way that the resulting model is compatible with optimization of predetermined classes of functions defined on those objects. We demonstrate the idea with low-dimensional models of policy matrices which contain all deterministic policies and show the efficiency of natural gradient methods within a reward maximization setting. Our construction can be used to optimize any linear program within a two-dimensional search space. This works particularly well if the graph of the *feasible region* (the polytope defined through the linear inequality constraints) allows a Hamiltonian cycle, as is the case for the *assignment problem* and the Birkhoff polytope of doubly stochastic matrices.

This thesis collects and extends the author's work [14, 86, 85, 87, 88]. Numerical experiments comprise custom implementation of a variety of algorithms; including Contrastive Divergence

methods, Expectation-Maximization algorithms, ordinary and natural-gradient parameter updates.

Notation

In the following we introduce the basic notation used in this work. In the various chapters we introduce more specific concepts. For example, exponential families are defined in Chapter 1 and Restricted Boltzmann Machines in Chapter 3.

We consider vectors of discrete and finite valued random variables. If X is a random variable taking values on a set $\mathcal{X}$, then $\mathcal{X}$ is called the *state space* (*sample space*) of X. A collection of $n \in \mathbb{N}$ random variables X_i with state spaces $\mathcal{X}_i$ for $i \in [n] := \{1, \dots, n\}$ is gathered into a random vector $X = (X_1, \dots, X_n)$ with state vectors $(x_1, \dots, x_n)$ in the product state space $\mathcal{X} := \times_{i=1}^n \mathcal{X}_i$. We usually denote x a state of the variable X. We call the variable X_i a q-ary variable when $\mathcal{X}_i$ has cardinality $q \in \mathbb{N}$. In this case $\mathcal{X}_i$ can be identified with the set $\{1, \dots, q\}$. Sometimes we find it more convenient to identify $\mathcal{X}_i$ with the set $\mathbb{F}_q = \mathbb{Z}/q\mathbb{Z} = \{0, \dots, q-1\}$ endowed with its algebraic structure. Given some $\mathcal{X} = \times_{i \in [n]} \mathcal{X}_i$ and any $\lambda \subseteq [n]$ we denote x_λ an element of $\times_{i \in \lambda} \mathcal{X}_i$ or the natural restriction of some $x \in \mathcal{X}$ to the coordinates $i \in \lambda$. If $\lambda = \{i, i+1, \dots, i+k\}$ we also write x_i^{i+k} for x_λ. The expression $[x_\lambda]$ represents a *cylinder set* of dimension $(n - |\lambda|)$. It consists of all $y \in \mathcal{X}$ with $y_\lambda = x_\lambda$. In the case of n binary variables, the cylinder sets are in one-to-one correspondence to the sets of vertices incident to faces of the n-dimensional unit cube $\{(r_1, \dots, r_n) \in \mathbb{R}^n \colon 0 \leq r_i \leq 1 \ \forall i\}$ and we also call them *cubical sets* or *faces of the unit n-cube*, which is justified from the combinatorial point of view. The lattice of cubical sets in $\{0,1\}^n$ is denoted C_n.

The set of real valued functions on $\mathcal{X}$ is denoted $\mathbb{R}^\mathcal{X}$. An element $f \in \mathbb{R}^x$ is a vector with entries $f(x) \in \mathbb{R}$ for all $x \in \mathcal{X}$. For any $\mathcal{Y} \subseteq \mathcal{X}$ we denote $\mathbb{1}_\mathcal{Y}$ the indicator function defined by $\mathbb{1}_\mathcal{Y}(x) = 1$ if $x \in \mathcal{Y}$ and $\mathbb{1}_\mathcal{Y}(x) = 0$ otherwise. The constant unity function $\mathbb{1}_\mathcal{X}$ is abbreviated by $\mathbb{1}$. The support of a function $f \in \mathbb{R}^\mathcal{X}$ is the set $\operatorname{supp}(f) := \{x \in \mathcal{X} \colon f(x) \neq 0\}$.

A probability distribution on $\mathcal{X}$ is an element $p \in \mathbb{R}^\mathcal{X}$ with $p(x) \geq 0$ for all $x \in \mathcal{X}$ and $\sum_{x \in \mathcal{X}} p(x) = 1$. A *point measure* (Dirac delta distribution) δ_x is a probability distribution with $\delta_x(x) = 1$. For any $\mathcal{Y} \subseteq \mathcal{X}$ we denote $u_\mathcal{Y}$ the uniform distribution on $\mathcal{Y}$, i.e., $u_\mathcal{Y} = \mathbb{1}_\mathcal{Y}/|\mathcal{Y}|$. The set of all probability distributions on $\mathcal{X}$ is denoted $\overline{\mathcal{P}}(\mathcal{X})$ or just $\overline{\mathcal{P}}$ if $\mathcal{X}$ is clear. We also consider the set of strictly positive probability distributions

$$\mathcal{P}(\mathcal{X}) := \{p \in \mathbb{R}^\mathcal{X} \colon p(x) > 0, \sum_{x \in \mathcal{X}} p(x) = 1\} \ .$$

$\overline{\mathcal{P}}$ is the closure of $\mathcal{P}$ in the topology of $\mathbb{R}^\mathcal{X}$. A probability distribution $p \in \overline{\mathcal{P}}(\mathcal{X})$ has *full support* if $p(x) > 0$ for all $x \in \mathcal{X}$. The set $\overline{\mathcal{P}}(\mathcal{X})$ is a $(|\mathcal{X}| - 1)$-dimensional simplex (the convex hull of the point measures δ_x with $x \in \mathcal{X}$), and hence it is also called a *probability simplex*. A *polytope* is the convex hull of a finite number of points in $\mathbb{R}^d$. A *simplex* is a d-dimensional polytope which is the convex hull of $(d+1)$ points. A probability simplex can be written as the disjoint union of smaller (relatively open) probability simplices as $\overline{\mathcal{P}}(\mathcal{X}) = \cup_{\mathcal{Y} \subseteq \mathcal{X}, \mathcal{Y} \neq \emptyset} \mathcal{P}(\mathcal{Y})$. We abbreviate $\mathcal{P}(\{0,1\}^n)$ by $\mathcal{P}_n$ or, when more clarity is convenient, by $\mathcal{P}_{n,\text{bin}}$.

Given a probability distribution p_n of n variables, the *marginal distribution* p_k of a subset of the variables $\{1,\ldots,k\} \subseteq \{1,\ldots,n\}$ is defined by

$$p_k(x_1,\ldots,x_k) = \sum_{\substack{y \in \times_{i\in[n]}\mathcal{X}_i: \\ y_i = x_i\ \forall i \in [k]}} p_n(y_1,\ldots,y_n) \qquad \forall (x_1,\ldots,x_k) \in \times_{i\in[k]}\mathcal{X}_i\ .$$

Multivariate distributions are sometimes conveniently written as an n-way table with entries $p_{x_1,\ldots,x_n} = p(x_1,\ldots,x_n)$.

A *statistical model* $\mathcal{M}$ is just a subset of $\mathcal{P}$. Usually $\mathcal{M}$ is endowed with some further structure. We consider mostly parametric models for which we are given some *parameter space* Θ (usually equal to $\mathbb{R}^d$ for some $d \in \mathbb{N}$) and a map $\Theta \to \mathcal{P}; \theta \mapsto p_\theta$ with $\mathcal{M} = \{p_\theta \in \mathcal{P} \colon \theta \in \Theta\}$.

There are several notions of distance between probability distributions and, in turn, for the error in the representation (approximation) of a probability distribution. One may use the induced distance of the Euclidian space of real valued functions $\mathbb{R}^{\mathcal{X}}$. However, from the point of view of information theory, a more meaningful distance notion for probability distributions is the *Kullback-Leibler divergence* (KL-divergence):

$$D(p\|q) := \sum_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{q(x)}\ .$$

We use the basis-2 logarithm. The KL-divergence is non-negative and vanishes if and only if $p = q$. If the support of q does not contain the support of p it is defined as ∞. The summands with $p(x) = 0$ are set to 0. The KL-divergence is not symmetric but it has nice information theoretic properties (see [71, 29]).

The Hamming distance between two vectors $x, y \in \times_{i\in[n]}\mathcal{X}_i$ is defined as $d_H(x,y) := |\{i \in [n] \colon x_i \neq y_i\}|$. An n-bit *binary code* is just a subset of $\{0,1\}^n$. We denote by $Z_{+,n}$ (or just Z_+ if n is clear) the binary code of length n consisting of vectors with an even number of entries equal to one:

$$Z_{+,n} := \left\{x \in \mathcal{X} = \{0,1\}^n \colon \sum_{i\in[n]} x_i = 0 \mod 2\right\}\ .$$

Similarly $Z_{-,n}$ denotes the set of binary vectors with an odd number of entries equal to one. The distance between any pair $x, y \in Z_\pm$ satisfies $d_H(x,y) = 0 \mod 2$. We find it sometimes useful to write the condition on the parity of a vector x as $\prod_{i\in[n]}(-1)^{x_i} = \pm 1$.

Given any subset $\mathcal{M}$ of the Euclidian space $\mathbb{R}^d$, the affine hull $\mathrm{aff}(\mathcal{M})$ is the smallest affine space in $\mathbb{R}^d$ that contains $\mathcal{M}$. The convex hull $\mathrm{conv}(\mathcal{M})$ consists of all convex combinations of finite sets of points from $\mathcal{M}$, i.e., points of the form $\sum_{i=1}^k \alpha_i p_i$ for some $k \in \mathbb{N}$, $\sum_{i=1}^k \alpha_i = 1$, $\alpha_i \geq 0$ and $p_i \in \mathcal{M}$ for all $i \in [k]$.

The expression "w.l.o.g." is an abbreviation for "without loss of generality", "s.t." abbreviates "such that", and "iff" stands for "if and only if".

Part I.
Exponential Families and Mixture Models

1 Mixtures of Discrete Exponential Families

In this chapter we use the combinatorics of support sets of distributions contained in the closures of discrete exponential families to assess the expressive power of their mixture models. This combinatorial approach appears natural in the light of recent advances on boundaries of exponential families [64, 65, 66, 96] and the algebraic perspective on graphical models proposed in [48].

Given a statistical model $\mathcal{M} \subseteq \mathcal{P}$ and a natural number $m \in \mathbb{N}$, the m-mixture of $\mathcal{M}$ is the set of probability distributions that can be written as the convex combination of at most m distributions from $\mathcal{M}$:

$$\mathrm{Mixt}^m(\mathcal{M}) := \Big\{ \sum_{j=1}^{m} \alpha(j) f_j : f_j \in \mathcal{M}, \alpha(j) \geq 0 \ \ \forall j \text{ and } \textstyle\sum_{j=1}^{m} \alpha(j) = 1 \Big\} .$$

The numbers $\alpha(j) \in \mathbb{R}$ are called *mixture weights* and the summands f_j *mixture components*. In this chapter we search for the smallest m for which $\mathrm{Mixt}^m(\mathcal{E}) \supseteq \mathcal{E}'$ for two exponential families $\mathcal{E}$ and $\mathcal{E}' \subseteq \mathrm{conv}(\mathcal{E})$.

Section 1.1 fixes notation and reviews basic facts about exponential families, their support sets and convex support polytopes. In Section 1.2 we discuss the support sets of faces of the probability simplex which are entirely contained in the closure of statistical models. We refer to these sets as S-sets of the statistical model. Section 1.3 analyzes coverings of the state space of hierarchical models using S-sets and contains the main results of this chapter; a description of necessary and sufficient number of mixture components from independence models and binary k-interaction models to represent arbitrary probability distributions or larger k-interaction models. Some proofs and details are shifted to Appendix 1.A, in order to improve readability of the main results. In Appendix 1.B we elaborate on the idea that the support sets of distributions within a mixture model provide information about the modes (topography) of distributions within the mixture model. We use this idea to bound the volume of complements of mixture models of binary independence models. Appendix 1.C analyzes submatrices of Hadamard matrices and properties of the support sets of exponential families that can be defined using them.

1.1 Exponential Families

There is an abundant literature on exponential families, see [10, 42, 23] for reference works. This section introduces notation and important concepts for the formulation of our results in the next sections.

Definition 1.1.1. Given a strictly positive function $\nu \in \mathbb{R}^{\mathcal{X}}_{>}$ and a linear subspace $V \subseteq \mathbb{R}^{\mathcal{X}}$, we define an *exponential family* $\mathcal{E}_{\nu,V}$ on $\mathcal{X}$ as the image of the following map:

$$\exp_\nu : \quad V \to \mathcal{P}(\mathcal{X}) ; \quad f \mapsto \nu \exp(f) / \sum_{x \in \mathcal{X}} \nu(x) \exp(f(x)) .$$

In particular an exponential family is a manifold. The differential geometry of exponential families has been studied within information geometry, see [6] for a reference text. Within algebraic statistics exponential families are described by toric varieties and are called toric-models (when V is generated by integer-valued functions), see [48, 109]. Exponential families are also referred to as log-linear models.

The function ν is called a *reference measure* of $\mathcal{E}_{\nu,V}$. There is no loss of generality in assuming that ν has full support $\mathcal{X}$. It is well known that $\mathcal{E}_{\nu,V} = \mathcal{E}_{\nu',V'}$ if and only if $\frac{\nu'}{\sum_x \nu'(x)} \in \mathcal{E}_{\nu,V}$ and $V' = V \mod \mathbb{1}$. The results from this chapter are independent of the the chosen reference measures. We refer to the space $\mathcal{T} = (V + \mathbb{R}\mathbb{1})/\mathbb{1}$ as the *tangent space* of $\mathcal{E}_{\nu,V}$, following the denotation proposed by Rauh [95]. There exists an isomorphism between $\mathcal{T}$ and $\mathscr{T}_p(\mathcal{E}_{\nu,V})$, the *differential geometric tangent space* at any point $p \in \mathcal{E}_{\nu,V}$.

A matrix $A \in \mathbb{R}^{d \times \mathcal{X}}$ with row span $\mathbb{R}^d \cdot A = V$ is called a *sufficient statistics* of $\mathcal{E}_{\nu,V}$. The rows of A, denoted $A_1, A_2, \ldots, A_d$, are functions on $\mathcal{X}$ called *observables*. We write A_x for the column vector $(A_1(x), A_2(x), \ldots, A_d(x))^\top$ for any $x \in \mathcal{X}$. We use subscripts $x \in \mathcal{X}$ to signify that A_x is a column, and subscripts $i \in [d]$ to signify that A_i is a row. The probability distributions in $\mathcal{E}_{\nu,V}$ can be given as:

$$p_\theta(x) = \nu(x) \exp(\theta^\top A_x - \psi_\theta) \quad \forall x \in \mathcal{X} \quad \forall \theta \in \mathbb{R}^d \,, \tag{1.1}$$

where $\psi_\theta := \log(\sum_y \nu(y) \exp(\theta^\top A_y))$ ensures that the entries of p_θ add up to one. The vector θ is called the *natural parameter* of p_θ. For convenience we always denote a sufficient statistics by A and we write $\mathcal{E}_{\nu,A}$ and $\mathcal{E}_{\nu,V}$ interchangeably if A is a sufficient statistics of $\mathcal{E}_{\nu,V}$. Furthermore, we omit the subscripts ν, V, A and write just $\mathcal{E}$ whenever there is no risk of confusion. The parametrization given in eq. (1.1) depends on A, but the exponential family only depends on $(V + \mathbb{R}\mathbb{1})/\mathbb{1}$. The map $\theta \mapsto p_\theta$ is one to one and $\mathcal{E}$ has dimension d if and only if $\{A_1, \ldots, A_d, \mathbb{1}\}$ are linearly independent (see [16, 10]). There is no loss of generality in including $\mathbb{1}$ as an observable, $A_{d+1} = \mathbb{1}$, and we do so throughout our considerations.

The probability distributions contained in an exponential family are strictly positive; they are contained in the open simplex $\mathcal{P}$. An exponential family on $\mathcal{X}$ approaches the boundary of $\mathcal{P}(\mathcal{X})$ at particular subsets $\mathcal{P}(\mathcal{Y}) \subset \partial\mathcal{P}(\mathcal{X})$ with $\mathcal{Y} \subset \mathcal{X}$. We will discuss them further below. The topological closure of $\mathcal{E}$ in $\mathbb{R}^{\mathcal{X}}$ is denoted $\overline{\mathcal{E}}$. In general this contains probability distributions which are not strictly positive. We call $\partial\mathcal{E} := \overline{\mathcal{E}} \setminus \mathcal{E}$ the *boundary* of $\mathcal{E}$.

In the multivariate case $X = (X_1, \ldots, X_n)$ the state space $\mathcal{X}$ is a Cartesian product $\bigtimes_{i \in [n]} \mathcal{X}_i$ and $x = (x_1, \ldots, x_n)$. A common choice of the spaces V defining $\mathcal{E}_V$ are spaces generated by functions of a limited number of variables. Given a collection of sets $\Delta \subseteq 2^{[n]}$ we consider the following linear space of functions:

$$V_\Delta := \Big\{ \sum_{\lambda \in \Delta} f_\lambda \in \mathbb{R}^{\mathcal{X}} \colon f_\lambda(x_\lambda, x_{[n]\setminus\lambda}) = f_\lambda(x_\lambda, \tilde{x}_{[n]\setminus\lambda}) \;\; \forall x, \tilde{x} \in \mathcal{X}, \; \forall \lambda \in \Delta \Big\} \,. \tag{1.2}$$

The functions f_λ only depend on the variables X_i with $i \in \lambda$ and the functions $f_\emptyset$ are constant. The collection Δ is called a *simplicial complex* on $[n]$ when it is inclusion complete, i.e., $\lambda \in \Delta$ implies $\lambda' \in \Delta$ for every $\lambda' \subseteq \lambda$, and furthermore $\{1, \ldots, n\} \subseteq \cup_{\lambda \in \Delta} \lambda$.

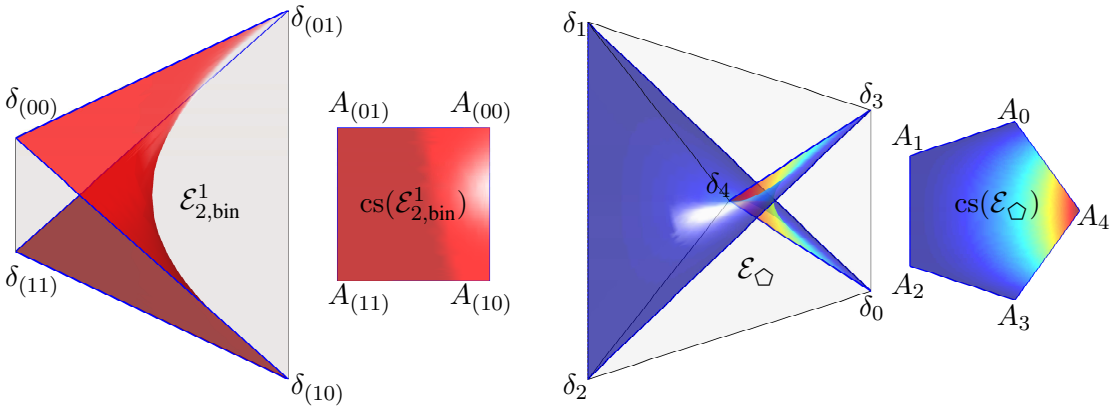

Figure 1.1: *Left: The three-dimensional simplex of probability distributions on* $\{0,1\}^2$*, the set of product distributions* $\mathcal{E}^1_{2,\text{bin}}$ *and its convex support* $\text{cs}(\mathcal{E}^1_{2,\text{bin}})$*. Right: Schlegel diagram*[1] *of the four dimensional probability simplex on* $\{0,1,\ldots,4\}$*, the corresponding projection of a two-dimensional exponential family* $\mathcal{E}_{\pentagon}$ *with convex support* $\text{cs}(\mathcal{E}_{\pentagon})$*. The color indicates the value that the distributions take on* $x = 4$*; blue for* $p(4) = 0$ *and red for* $p(4) = 1$*. The uniform distribution* $\frac{\mathbb{1}}{5}$ *as well as* δ_4 *are projected into the same point.*

Definition 1.1.2. If Δ is a simplicial complex on $[n]$, then we call $\mathcal{E}_\Delta := \mathcal{E}_{V_\Delta}$ the *hierarchical model* of X with interaction structure Δ. Given any natural number $k \leq n$, the model $\mathcal{E}^k := \mathcal{E}_{V_{\Delta_k}}$ with $\Delta_k := \{\lambda \subseteq [n] : |\lambda| \leq k\}$ is called *k-interaction model*. The important special case $\mathcal{E}^1$ consists of product distributions and is called *independence model*.

Any k-interaction model is a hierarchical model. There is a natural hierarchy of nested models $\mathcal{E}^1 \subset \mathcal{E}^2 \subset \cdots \subset \mathcal{E}^n = \mathcal{P}$. See [6] for details on these hierarchies. We write $\mathcal{E}^k_{n,\text{bin}}$ for the k-interaction model of n binary variables, and $\mathcal{E}^k_{n,q\text{-ary}}$ for the k-interaction model of n q-ary variables. The dimension of the hierarchical model $\mathcal{E}_\Delta$ on $\bigtimes_{i=1}^n \mathcal{X}_i$ is $\dim(\mathcal{E}_\Delta) = \sum_{\lambda\in\Delta}\prod_{i\in\lambda}(|\mathcal{X}_i|-1)-1$ (see [61]). In particular, the binary k-interaction model has dimension $\dim(\mathcal{E}^k_{n,\text{bin}}) = \sum_{i=1}^k \binom{n}{i}$. More details on hierarchical models can be found in [66, 61].

The independence model $\mathcal{E}^1$ consists of probability distributions of the form $p(x_1,\ldots,x_n) = \exp(\sum_{i\in[n]} f_i(x_i)) = p_1(x_1)\cdots p_n(x_n)$, with $p_i \in \mathcal{P}(\mathcal{X}_i)$ for all $i \in [n]$. Any probability distribution p on $\mathcal{X}$ can be written conveniently as an n-way $(|\mathcal{X}_1| \times \cdots \times |\mathcal{X}_n|)$-table (tensor) with entries $p_{x_1,\ldots,x_n} = p(x_1,\ldots,x_n)$ for all $x \in \mathcal{X}$. The probability distributions in $\overline{\mathcal{E}^1}$ correspond to the tables $p_1 \otimes p_2 \otimes \cdots \otimes p_n$, where each $p_i \in \overline{\mathcal{P}(\mathcal{X}_i)}$ is a non-negative real vector of length $|\mathcal{X}_i|$, and correspond, up to normalization, to the set of non-negative $(|\mathcal{X}_1| \times \cdots \times |\mathcal{X}_n|)$-tensors of rank one.

In algebraic geometry, the independence model is considered as a subset of the *Segre embedding* of a product of projective spaces $\mathbb{P}^{|\mathcal{X}_i|-1}$, which is the map defined through $\mathbb{P}^{q_1-1} \times \mathbb{P}^{q_2-1} \to \mathbb{P}^{q_1q_2-1}$ and $((X_1 : \cdots : X_{q_1}),(Y_1 : \cdots : Y_{q_2})) \mapsto (X_1Y_1 : X_1Y_2 : \cdots : X_{q_1}Y_{q_2})$, where $(X_1 : \cdots : X_q)$ denotes the equivalence class of $(X_1,\ldots,X_q)$ in projective space (called

[1] A *Schlegel diagram* is a projection of a polytope onto the affine hull of one of its facets by rays through a point *beyond* that facet. It yields a *complex* which is combinatorially equivalent to the original polytope. See [51] for details.

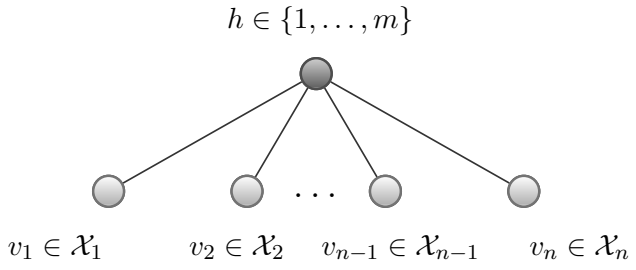

Figure 1.2: *Graphical representation of the m-mixture of the independence model of n variables v_i, $i \in [n]$. Each node in the bottom represents one of the variables v_i. The dark node represents a hidden variable with m states. If $\mathcal{X}_i = \{0, \ldots, q-1\}$ for all i, this is the model* $\mathrm{Mixt}^m(\mathcal{E}^1_{n,q\text{-ary}})$.

homogeneous coordinates). Figure 1.1 left shows the model $\mathcal{E}^1_{2,\text{bin}}$. We will use this example to illustrate various concepts in the remainder of this section.

Remark 1.1.3. Any distribution in the m-mixture of the independence model of n variables has the form $p(x_1, \ldots, x_n) = \sum_{j\in[m]} \alpha(j) \exp(\sum_{i\in[n]} f_i^j(x_i))$, where f_i^j is a function of the variable x_i only. Regarding j as the state of a hidden variable X_0, we can write this as

$$p(x_1, \ldots, x_n) = \sum_{x_0\in[m]} \exp(f_0(x_0)) \exp(\sum_{i\in[n]} f_{\{i,0\}}(x_i, x_0)) \; . \tag{1.3}$$

Hence the m-mixture of the independence model is the marginal of a hierarchical model with maximal interaction sets $\{\{1,0\}, \ldots, \{n,0\}\}$. This model has the graphical representation depicted in Figure 1.2. Each edge in the graphical representation corresponds to an interaction set of cardinality two.

A Hadamard matrix is a ± 1 matrix with orthogonal rows. A natural choice for the sufficient statistics of binary hierarchical models are submatrices of the following *Hadamard matrix*:

$$A = (A_{\lambda,x})_{\lambda\in 2^{[n]}, x\in\mathcal{X}} \; , \; A_{\lambda,x} := (-1)^{|\mathrm{supp}(x)\cap\lambda|} \; . \tag{1.4}$$

In this case the observables are well studied functions referred to as *characters*, and if $\Delta \subseteq 2^{[n]}$ is inclusion complete, then the rows A_λ with $\lambda \in \Delta$ build an orthogonal basis of V_Δ. In Appendix 1.C we discuss Hadamard matrices in more detail.

Example 1.1.4. The binary independence model $\mathcal{E}^1_{n,\text{bin}}$ has a sufficient statistics with columns $A_x = ((-1)^{x_i})_{i\in[n]} = (-2x + \mathbb{1})$ for $x \in \{0,1\}^n$. In particular, the set of columns $\{A_x\}_x$ is the list of vectors in $\{+1,-1\}^n$ and $\mathrm{conv}\{A_x\}$ is an n-dimensional hypercube. Furthermore, any binary product distribution can be written as $p(x) = \exp(\langle\theta, x\rangle - \psi_\theta)$, where $\theta \in \mathbb{R}^n$, $\langle\theta, x\rangle = \sum_{i\in[n]} \theta_i x_i$, and ψ_θ ensures normalization.

We will need polytopes and some related notions summarized in the following definition:

Definition 1.1.5. A *polytope* is a compact convex subset of $\mathbb{R}^d$ with finitely many extreme points. Let $Q \subset \mathbb{R}^d$ be a polytope. A proper *face* of Q is the intersection of Q with a hyperplane of codimension one in $\mathbb{R}^d$ such that all points of Q lie on one of the closed half-spaces defined through that hyperplane. The polytope itself is a face (called improper face). The set of faces is denoted $\mathcal{F}(Q)$. The dimension of a face F is $\dim(F) := \dim \mathrm{aff}(F)$. A *facet* is a proper face

of maximal dimension. An extreme point is a zero-dimensional face. The set of extreme points is denoted $\mathrm{ex}(Q)$. The *combinatorial type* of Q is the set of all faces $\mathcal{F}(Q)$ together with the partial order of inclusion relations. The polytope $Q \subset \mathbb{R}^d$ is called *K-neighborly* if the convex hull of any K or less of its vertices is a face (see [67, 105]).

Remark 1.1.6. If Q is a polytope and $0 \leq g \leq \dim(Q) - 1$, then the union of g-dimensional faces $\cup_{F \in \mathcal{F}(Q):\dim(F)=g} F$ contains all vertices of Q (see [54, Theorem 15.1.2]). Furthermore, any nonsingular affine transformation of a polytope yields a combinatorially equivalent polytope (see [51, Theorem 3.2.3]).

Definition 1.1.7. A d-dimensional cyclic polytope with v vertices $C(v, d)$ is defined as the convex hull of v different points on the *d-moment curve*: $C(v, d) := \mathrm{conv}\{f(t_i)\}_{i \in I, |I|=v}$, where $v \geq d + 1$, I is a linearly ordered set, $i \mapsto t_i$ is a strictly monotone function, and f is $\mathbb{R} \to \mathbb{R}^d\,;\ t \mapsto (t, t^2, \ldots, t^d)$. See [46].

Definition 1.1.8. The *convex support* of $\mathcal{E}_A$ is a convex polytope $\mathrm{cs}(\mathcal{E}_A)$ with the following vertex presentation (i.e., as convex hull of a set of points):

$$\mathrm{cs}(\mathcal{E}_A) := \mathrm{conv}\{A_x\}_{x \in \mathcal{X}} \subset \mathbb{R}^d\ .$$

Note that not every A_x must be an extreme point of $\mathrm{cs}(\mathcal{E}_A)$. The combinatorial type of the polytope $\mathrm{cs}(\mathcal{E}_A)$ is determined by V, the row span of A. In the case of hierarchical models, the convex support is also called *marginal polytope*.

The sufficient statistics matrix A induces the *moment map*: $\overline{\mathcal{P}} \to \mathbb{R}^d\,;\ p \mapsto A \cdot p$, which maps $\overline{\mathcal{E}_A}$ bijectively onto $\mathrm{cs}(\mathcal{E}_A)$ (see [23]). This bijective map is in fact a homeomorphism, because the moment map is continuous, $\overline{\mathcal{E}_A}$ is compact, and $\mathrm{cs}(\mathcal{E}_A)$ is Hausdorff. The vector $A \cdot p$ contains the p-expectation values of the observables and is called the *expectation parameter* of p. For any $\eta \in \mathrm{cs}(\mathcal{E}_A)$ we denote p_η the unique probability distribution in $\overline{\mathcal{E}_A}$ with $A \cdot p_\eta = \eta$.

Definition 1.1.9. A set $\mathcal{Y} \subseteq \mathcal{X}$, $\mathcal{Y} \neq \emptyset$ is a *facial* set of $\mathcal{E}_A$ iff $\mathcal{Y} = \{x \in \mathcal{X} \colon A_x \in F\}$ for some face F of $\mathrm{cs}(\mathcal{E}_A)$. We denote the set of facial sets of $\mathcal{E}$ by $\mathcal{F}(\mathcal{E})$.

A starting point for the ideas developed in this chapter is the following well known fact (see [48, 96] for example):

Lemma 1.1.10. *A set $\mathcal{Y}$ is the support set of some distribution $p \in \overline{\mathcal{E}}$ if and only if $\mathcal{Y}$ is facial.*

The map $G \in \mathcal{F}(\mathrm{cs}(\mathcal{E}_A)) \mapsto \{x \in \mathcal{X} \colon A_x \in G\} =: \mathcal{X}_G \in \mathcal{F}(\mathcal{E}_A)$ is an isomorphism between the face lattice of $\mathrm{cs}(\mathcal{E}_A)$ and the support sets of distributions within $\overline{\mathcal{E}_A}$ which preserves the partial order of inclusion.

Example 1.1.11. The facial sets of $\mathcal{E}^1_{n,\mathrm{bin}}$ are the sets of vertices incident to faces of the n-dimensional unit cube, which are precisely the cylinder sets of $\{0,1\}^n$. In the case $n = 2$ the independence model has a sufficient statistics with columns $A_{00} = (-1,-1)^\top$, $A_{01} = (1,-1)^\top$, $A_{10} = (-1,1)^\top$, $A_{11} = (1,1)^\top$. See Figure 1.1. The facial sets are $\mathcal{X}$, the pairs $\{(0,0),(0,1)\}$, $\{(0,1),(1,1)\}$, $\{(1,1),(1,0)\}$, $\{(1,0),(0,0)\}$, which correspond to edges of the convex support, and the individual elements of $\mathcal{X}$, which correspond to the vertices of the convex support. The edges and vertices of $\mathrm{cs}(\mathcal{E}^1_{2,\mathrm{bin}})$ are the only simplex faces.

1.2 *S*-sets

We introduce the following type of support sets of a model:

Definition 1.2.1. Given a set of probability distributions $\mathcal{M} \subseteq \overline{\mathcal{P}}(\mathcal{X})$ we say that a set $\mathcal{Y} \subseteq \mathcal{X}$ is an *S-set* of $\mathcal{M}$ iff every distribution with support $\mathcal{Y}$ is contained in the closure $\overline{\mathcal{M}}$.

In the following we work out properties of S-sets and relate the combinatorics of support sets to the expressive power of mixtures of exponential families.

Given an exponential family $\mathcal{E}$ on $\mathcal{X}$ we consider the following function which gives the minimal cardinality of a facial *packing* of any set $\mathcal{Z} \subseteq \mathcal{X}$:

$$\kappa^f_{\mathcal{E}} : \ 2^{\mathcal{X}} \to \mathbb{N} \ ; \ \mathcal{Z} \mapsto \min\{n \in \mathbb{N} \colon \exists\, \mathcal{Y}_1, \ldots, \mathcal{Y}_n \in \mathcal{F}(\mathcal{E}) \text{ with } \cup_i \mathcal{Y}_i = \mathcal{Z}\} \ .$$

We set $\kappa^f_{\mathcal{E}}(\mathcal{Z}) = \infty$ if there doesn't exist a facial packing of $\mathcal{Z}$. We write κ^f_k for $\kappa^f_{\mathcal{E}_k}$. All $\mathcal{Y}_i$ in a packing of $\mathcal{Z}$ are subsets of $\mathcal{Z}$. If $\mathcal{E}_\Delta$ is a hierarchical model, then every $\{x\}$ is facial, (provided that $\cup_{\lambda\in\Delta}\lambda = [n]$), and $\kappa^f_{\mathcal{E}_\Delta} < \infty$.

We consider the smallest number of S-sets that cover any $\mathcal{Z} \subseteq \mathcal{X}$, which is the following function:

$$\kappa^s_{\mathcal{E}} : \ 2^{\mathcal{X}} \to \mathbb{N} \ ; \ \mathcal{Z} \mapsto \min\{n \in \mathbb{N} \colon \exists\, \mathcal{Y}_1, \ldots, \mathcal{Y}_n \ S\text{-sets with } \cup_i \mathcal{Y}_i \supseteq \mathcal{Z}\} \ .$$

We set $\kappa^s_{\mathcal{E}}(\mathcal{Z}) = \infty$ if there doesn't exist an S-set covering of $\mathcal{Z}$. All S-sets of $\mathcal{E}$ are contained in $\mathcal{F}(\mathcal{E})$. If κ S-sets cover $\mathcal{X}$, then at most κ S-sets are needed for packing any $\mathcal{Z} \subseteq \mathcal{X}$, because any subset of an S-set is an S-set. We abbreviate $\kappa^s_{\mathcal{E}}(\mathcal{X})$ with $\kappa^s_{\mathcal{E}}$. Given two exponential families $\mathcal{E}$ and $\mathcal{E}'$ we consider also the maximum of $\kappa^f_{\mathcal{E}}$ restricted to the facial sets of $\mathcal{E}'$:

$$\kappa^f_{\mathcal{E},\mathcal{E}'} := \max_{\mathcal{Z}\in\mathcal{F}(\mathcal{E}')} \kappa^f_{\mathcal{E}}(\mathcal{Z}) \ . \tag{1.5}$$

The functions $\kappa^s_{\mathcal{E}}$ and $\kappa^f_{\mathcal{E}}$ can be easily defined for arbitrary models $\mathcal{M} \subseteq \overline{\mathcal{P}}$ instead of $\mathcal{E}$, replacing "*facial sets*" by "*support sets of distributions within $\overline{\mathcal{M}}$*" in the above definitions.

We can find sufficient and necessary numbers of mixture components to represent a distribution $q \in \mathcal{P}$ with support $\mathrm{supp}(q) = \mathcal{Y}$ by deriving bounds on the number of S-sets which is sufficient to cover $\mathcal{Y}$, and on the number of facial sets needed to pack $\mathcal{Y}$. The following lemma describes this very natural observation:

Lemma 1.2.2. *Consider two exponential families $\mathcal{E}, \mathcal{E}' \subseteq \mathcal{P}(\mathcal{X})$.*

- *If $m \geq \kappa^s_{\mathcal{E}} < \infty$, then $\mathrm{Mixt}^m(\mathcal{E}) = \mathcal{P}$.*
- *$\mathrm{Mixt}^m(\overline{\mathcal{E}}) \supseteq \overline{\mathcal{E}'}$ implies $m \geq \kappa^f_{\mathcal{E},\mathcal{E}'}$.*

In particular, $\mathrm{Mixt}^m(\overline{\mathcal{E}}) = \overline{\mathcal{P}}$ implies $m \geq \max \kappa^f_{\mathcal{E}}$, and $\kappa^f_{\mathcal{E},\mathcal{E}'} = \infty$ implies $\mathrm{conv}(\mathcal{E}) \not\supset \mathcal{E}'$. This lemma can be formulated for arbitrary models. In that case however, the implication of the first item holds only for the closures: If $m \geq \kappa^s_{\mathcal{M}}$, then $\mathrm{Mixt}^m(\overline{\mathcal{M}}) = \overline{\mathcal{P}}$.

Proof of Lemma 1.2.2. 1. Let $\{\mathcal{Y}_i\}_{i=1}^{\kappa}$ be an S-set covering of $\mathcal{X}$. W.l.o.g. $\mathcal{Y}_i \cap \mathcal{Y}_j = \emptyset \; \forall i \neq j$. Any $p \in \overline{\mathcal{P}}$ can be written as $\sum_{i=1}^{\kappa} \alpha_i f_i$ and $f_i \in \overline{\mathcal{E}}$ choosing f_i with $\operatorname{supp}(f_i) \subseteq \mathcal{Y}_i$, $f_i = p|_{\mathcal{Y}_i} / \sum_{x \in \mathcal{Y}_i} p(x)$ and $\alpha_i = \sum_{x \in \mathcal{Y}_i} p(x)$. I.e., $\operatorname{Mixt}^{\kappa}(\overline{\mathcal{E}}) = \overline{\mathcal{P}}$. For strictly positive distributions: Clearly, $\operatorname{Mixt}^{m}(\mathcal{E}) \subseteq \mathcal{P}$. For the other direction: Let $Y_i := \overline{\mathcal{P}}(\mathcal{Y}_i)$. The sets Y_i are disjoint faces of $\overline{\mathcal{P}}$ whose union covers all point measures $\{\delta_x\}_{x \in \mathcal{X}}$. Let $p_\eta := (A|_{\overline{\mathcal{E}}})^{-1}(\eta)$ (the unique probability distribution in $\overline{\mathcal{E}}$ with expectation vector η). We have just seen that the mixture map $\phi : D := \overline{\mathcal{P}}_m \times (\times_{i=1}^{m} \operatorname{cs}(\mathcal{E})) \to \overline{\mathcal{P}} \; ; \; (\alpha, \eta_1, \ldots, \eta_m) \mapsto \sum_{i=1}^{m} \alpha(i) p_{\eta_i}$ is surjective. It is easy to check that the restriction $\phi|_C \colon C \to \partial\mathcal{P}$, with $C := \partial(\mathcal{P}_m \times (\times_{i=1}^{m}(A \cdot Y_i)))$ is a continuous bijection between the compact domain C and the Hausdorff codomain $\partial\mathcal{P}$. Therefore, $\phi|_C$ is a homeomorphism and induces isomorphisms between the homotopy groups of C and those of $\partial\mathcal{P} \simeq S^{|\mathcal{X}|-2}$ ($S^{|\mathcal{X}|-2}$ denotes the $(|\mathcal{X}|-2)$-sphere). Note that $\phi(\mathring{D}) \subseteq \mathcal{P}$. For any $\epsilon > 0$ we find a continuous deformation $C \to \tilde{C} \subseteq \mathring{D}$ which is mapped by ϕ into a continuous deformation $\partial\mathcal{P} \to \phi(\tilde{C}) \subset \mathcal{P} \setminus \mathcal{P}^{\epsilon}$, $\mathcal{P}^{\epsilon} := \{p \in \mathcal{P} \colon p(x) \geq \epsilon \; \forall x \in \mathcal{X}\}$. If $\phi(\mathring{D})$ didn't contain $\mathcal{P}^{\epsilon}$, then $\phi(\tilde{C})$ wouldn't be contractible in $\phi(\mathring{D})$, in contradiction to the fact that $\mathring{D}$ is contractible. Obviously any element of $\mathcal{P}$ belongs to some $\mathcal{P}^{\epsilon}$. Hence, $\operatorname{Mixt}^{m}(\mathcal{E}) \supseteq \mathcal{P}$.

2. Consider some $p \in \overline{\mathcal{E}'}$ with a support $\mathcal{Z} \in \mathcal{F}(\mathcal{E}')$. If p is written as a mixture of elements from $\overline{\mathcal{E}}$, then every summand with positive mixture weight must have a support $\mathcal{Y} \in \mathcal{F}(\mathcal{E})$ with $\mathcal{Y} \subseteq \mathcal{Z}$. Furthermore, the union of the support sets of these summands must be $\mathcal{Z}$. The minimal number is precisely $\kappa^{f}_{\mathcal{E}}(\mathcal{Z})$. □

Example 1.2.3. (Cylinder S-sets). Any distribution p with support contained in a cylinder set $[y_{\Lambda^c}] = \{x \in \mathcal{X} \colon x_{\Lambda^c} = y_{\Lambda^c}\}$, $\Lambda \subseteq \mathcal{X}$, $|\Lambda| = k$ is contained in $\overline{\mathcal{E}^k}$. Indeed, if $p \in \overline{\mathcal{P}}$ is arbitrary with support $[y_{\Lambda^c}]$, then $p(x) = \lim_{\alpha \to \infty} \exp(f(x_\Lambda) - \alpha \sum_{j \in \Lambda^c} g_j(x_j))/Z$, where Z is the normalization constant, $f(x) = f(x_\Lambda)$ is a function of k variables with $f(x_\Lambda) = \log(p(x)) + \log(Z) \; \forall x \in [y_{\Lambda^c}]$ and g_j are functions of one variable taking value 0 for $x_j = y_j$ and 1 otherwise. Therefore, the k-dimensional cylinder sets are S-sets of $\mathcal{E}^k$. If $\mathcal{X} = \{1, \ldots, q\}^n$, then $\kappa^{s}_{\mathcal{E}^k} \leq q^{n-k}$ and q^{n-k} mixtures of $\mathcal{E}^k$ suffice to represent any distribution. We will improve this bound in Theorem 1.3.9.

Example 1.2.4. (Mixtures of two independent binary variables). The set of mixtures of two fixed distributions p and q on $\mathcal{X}$ can be represented as a line segment in $\mathbb{R}^{\mathcal{X}}$ connecting the two points. If p and q are moved freely within $\overline{\mathcal{E}^1_{2,\text{bin}}}$, the set of segments fills the entire probability simplex $\overline{\mathcal{P}_2}$ (see Figure 1.1 left). Close inspection reveals that mixtures of elements from the intervals $[\delta_{(0,1)}, \delta_{(1,1)}]$ and $[\delta_{(1,0)}, \delta_{(0,0)}]$, already suffice to fill $\overline{\mathcal{P}_2}$. These intervals correspond to a partition of $\mathcal{X}$ into two S-sets of $\mathcal{E}^1_{2,\text{bin}}$.

The following lemma will help us estimate $\kappa^{s}_{\mathcal{E}}$ for specific choices of $\mathcal{E}$:

Lemma 1.2.5. (S-sets of exponential families). *Consider an exponential family $\mathcal{E}_A \subseteq \mathcal{P}(\mathcal{X})$. The following statements are equivalent:*

- *Every probability distribution with support $\mathcal{Y} \subseteq \mathcal{X}$ is in $\overline{\mathcal{E}}$, i.e., $\mathcal{Y}$ is an S-set.*
- *$\mathcal{Y}$ is facial and $\operatorname{conv}\{A_y\}_{y \in \mathcal{Y}}$ is a $(|\mathcal{Y}|-1)$-dimensional simplex.*
- *$\operatorname{supp}(m^{\pm}) \not\subset \mathcal{Y} \;\; \forall m \in \ker(A) \setminus \{0\} \subset \mathbb{R}^{\mathcal{X}}$, where $m^{\pm}(x) = \max\{0, \pm m(x)\} \;\; \forall x \in \mathcal{X}$.*

Proof. The first item implies the second because the linear map A is a bijection on the simplex $\mathcal{P}(\mathcal{Y})$. For the other direction: The matrix $A_{\mathcal{Y}} := (A_y)_{y \in \mathcal{Y}}$ defines an exponential family $\mathcal{E}_{\mathcal{Y}} = \overline{\mathcal{E}} \cap \mathcal{P}(\mathcal{Y})$, because $\mathcal{Y}$ is facial. If $\operatorname{conv}\{A_y\}_{y \in \mathcal{Y}}$ is a $(|\mathcal{Y}|-1)$-simplex, then all columns

of $A_{\mathcal{Y}}$ are affinely independent. In fact they are linearly independent ($\mathbb{1}$ is a row of A), and $\ker A_{\mathcal{Y}} = \{0\}$. In this case $\mathcal{E}_{\mathcal{Y}} = \mathcal{P}(\mathcal{Y})$. The third item is equivalent to: $\mathcal{Y}$ is facial [96] and additionally $\operatorname{supp}(m) \not\subset \mathcal{Y}\ \forall m \in \ker(A)$. This implies $\ker A_{\mathcal{Y}} = \{0\}$. □

In Appendix 1.A we provide a more extensive discussion of the results from Lemma 1.2.5.

Remark 1.2.6. By Lemma 1.2.5, if $|\operatorname{supp}(p)| < |\operatorname{supp}(m^+)|\ \forall m \in \ker(A) \setminus \{0\}$, then $p \in \overline{\mathcal{E}_A}$. Furthermore, there always exists some $q \in \overline{\mathcal{P}}(\mathcal{X}) \backslash \overline{\mathcal{E}_A}$ with $|\operatorname{supp}(q)| = \min_{m \in \ker(A)\setminus\{0\}} |\operatorname{supp}(m^+)|$.

If the convex support of an exponential family $\mathcal{E}$ is K-neighborly and $\overline{\mathcal{E}}$ contains all point measures, then $\overline{\mathcal{E}}$ contains any probability distribution with support of cardinality at most K. In other words, any $\mathcal{Y}$ with $|\mathcal{Y}| \leq K$ is an S-set.

Example 1.2.7. (An n-gon exponential family). Let $\mathcal{X} = \{0, \ldots, n-1\}$ and let $\mathcal{E}$ be an exponential family with convex support given by an n-gon (a convex polygon with n vertices). This is a two-dimensional family which contains all point measures δ_x in its closure. The later property finds applications in model design and will be the topic of Section 5.2. Assume that the boundary of $\operatorname{cs}(\mathcal{E})$ is given by the polyline $A_0A_1A_2 \cdots A_{n-1}A_0$. The facial sets are: $\mathcal{X}$, the pairs $\{i, i+1\mod n\}$ and the points $\{i\}_{i\in\mathcal{X}}$. All facial sets are S-sets, with exception of $\mathcal{X}$. The sample space $\mathcal{X}$ is covered by $\kappa^s_{\mathcal{E}} = \left\lceil \frac{n}{2} \right\rceil$ S-sets, while the packing of any set $\mathcal{Y} \subseteq \mathcal{X}$ requires at most $\max \kappa^f_{\mathcal{E}} = \left\lfloor \frac{n}{2} \right\rfloor$ facial sets. By Lemma 1.2.2 the smallest m for which $\operatorname{Mixt}^m(\mathcal{E}) = \operatorname{conv}(\mathcal{E}) = \mathcal{P}$ satisfies $\left\lfloor \frac{n}{2} \right\rfloor \leq m \leq \left\lceil \frac{n}{2} \right\rceil$. In the case $n = 5$ (see Figure 1.1 right) we can show that $m \geq 2 = \left\lfloor \frac{n}{2} \right\rfloor$ is necessary and sufficient:

Proposition 1.2.8. *Let $\mathcal{E}_{\pentagon}$ be an exponential family on $\{0, 1, \ldots, 4\}$ with convex support a pentagon, as the one shown in Figure 1.1 right. Then $\operatorname{Mixt}^2(\mathcal{E}_{\pentagon}) = \mathcal{P}$.*

Proof. See Appendix 1.A for the proof and a topological discussion. □

The following example will help us illustrate some results in the remainder of this chapter:

Example 1.2.9. (The convex support of $\mathcal{E}^2_{4,\text{bin}}$). The polytope $\operatorname{cs}(\mathcal{E}^2_{4,\text{bin}})$ has dimension 10 and 16 vertices. We used the computer software `Polymake` [47] to compute its face lattice. The polytope has 56 facets (proper faces of maximal dimension 9). From these, 16 contain only 10 vertices and are simplices. One of the corresponding S-sets is the following: $\mathcal{Y} = \{(0000), (1000), (0100), (0010), (1001), (0101), (0011), (1101), (1011), (0111)\}$. In total 8 S-sets contain each 6 elements from Z_+ and 8 contain 6 elements from Z_-. The other 40 facets have 12 vertices each. Denote $\{F_i\}$ the S-sets (of cardinality 10) and $\{G_i\}$ the remaining facets (of cardinality 12). We found that $F_i \cup F_j \neq \mathcal{X}\ \forall i, j$ and $F_i \cup G_j \neq \mathcal{X}\ \forall i, j$. Since all faces (facial sets) and in particular all simplex faces (S-sets) are subsets of some facet, these computations show that a minimal covering of $\mathcal{X}$ using S-sets has cardinality at least 3.

We briefly discuss symmetries of $\operatorname{cs}(\mathcal{E}_\Delta)$ with interesting relations to coding theory, and to the work [66]:

If the family of interaction sets Δ is invariant under permutations of the coordinate indices $\pi : [n] \to [n]$, then also $\operatorname{cs}(\mathcal{E}_\Delta)$. If $\mathcal{Y}$ is an S-set of $\mathcal{E}^k$, then $\pi(\mathcal{Y}) := \{(x_{\pi(1)}, \ldots, x_{\pi(n)}) : x \in \mathcal{Y}\}$ is also an S-set for any permutation π. A further symmetry is given by re-labeling the values of the variables:

Remark 1.2.10. Consider any $\Delta \subseteq 2^{[n]}$. If $\mathcal{E}_A$ is an exponential family with sufficient statistics $A = ((-1)^{|\operatorname{supp}(x)\cap\lambda|})_{\lambda\in\Delta, x\in\mathcal{X}}$, then $\mathcal{Y}$ is an S-set if and only if $x * \mathcal{Y} := \{x + y \mod 2 : y \in$

$\mathcal{Y}\}$ is an S-set for all $x \in \mathcal{X}$. Furthermore, $\mathcal{Y} \subseteq \mathcal{X}$ is facial if and only if $x * \mathcal{Y}$ is facial for all $x \in \mathcal{X}$.

The elements of the family $\{x * \mathcal{Y}\}_{x\in\mathcal{X}}$ need not be different from each other, but they are if $|\mathcal{Y}|$ is odd or if $\mathcal{Y}$ is a Hamming ball. For any $\mathcal{Y} \subseteq \mathcal{X}$, $\mathcal{Y} \neq \emptyset$ we have $\cup_{x\in\mathcal{X}} x * \mathcal{Y} = \mathcal{X}$, since $\{x * z : x \in \mathcal{X}\} = \mathcal{X}$ for any $z \in \mathcal{X}$. Moreover, $|x * \mathcal{Y}| = |\mathcal{Y}|$ for any $x \in \mathcal{X}$, $\mathcal{Y} \subseteq \mathcal{X}$, since $x * (x * \mathcal{Y}) = \mathcal{Y}$. In Appendix 1.C we provide more details on this.

1.3 Mixtures of Hierarchical Models

By Lemma 1.2.5 we can find $\kappa^s_{\mathcal{E}}$ if we determine the simplex faces of $\mathrm{cs}(\mathcal{E})$ and how many such faces suffice to cover all vertices of $\mathrm{cs}(\mathcal{E})$. This has the flavor of a covering code problem, which can be difficult; e.g., finding a minimum clique cover of a graph is a graph-theoretical NP-complete problem, and perfect covering codes on $\{0,1\}^n$ are not completely understood (see [27] for details on covering codes). Here we focus on S-set coverings and facial packings for hierarchical models.

Product Distributions

The set of strictly positive product distributions of n variables with state space $\mathcal{X} = \bigtimes_{i\in[n]} \mathcal{X}_i$ is:

$$\mathcal{E}^1 = \{p \in \mathcal{P} : p(x_1, \ldots, x_n) = \prod_{i\in[n]} p_i(x_i), p_i \in \mathcal{P}(\mathcal{X}_i)\} \; . \tag{1.6}$$

The convex support of $\mathcal{E}^1$ is a Cartesian product $\mathrm{cs}(\mathcal{E}^1) = \bigtimes_{i\in[n]} S_i$, where S_i is a $(|\mathcal{X}_i| - 1)$-dimensional simplex for every $i \in [n]$. The facial sets are:

$$\mathcal{F}(\mathcal{E}^1) = \{ \bigtimes_{i\in[n]} \mathcal{Y}_i : \mathcal{Y}_i \subseteq \mathcal{X}_i \; \forall i \in [n]\} \; . \tag{1.7}$$

The S-sets have the form $\{(x_1, \ldots, x_n) \in \mathcal{X} \colon x_i \in \mathcal{Y}_i$ and $x_j = y_j \; \forall j \neq i\}$ for some $i \in [n]$, $\mathcal{Y}_i \subseteq \mathcal{X}_i$, and some $y_j \in \mathcal{X}_j \; \forall j \neq i$. See [78] for interesting properties of products of simplices.

In the case of binary variables, the convex support $\mathrm{cs}(\mathcal{E}^1_{n,\mathrm{bin}})$ is a combinatorial n-cube. The set $\mathcal{Y} \subseteq \mathcal{X}$ supports a distribution in $\overline{\mathcal{E}^1}$ iff $\mathcal{Y}$ is a cylinder set, i.e., if there is some $\lambda \subseteq [n]$ and some y_λ with

$$\mathcal{Y} = \{(x_1, \ldots, x_n) \in \{0,1\}^n \colon x_i = y_i \; \forall i \in \lambda\} \; . \tag{1.8}$$

Hence $\mathcal{Y}$ is an S-set iff it has cardinality one or consists of two binary vectors with Hamming distance one, see Example 1.2.3.

We will use the following function in the formulation of the next theorem:

$$\mathcal{A}_q(n, d) := \max\{|\mathcal{Y}| : \mathcal{Y} \subseteq \mathcal{X} \text{ s.t. } d_H(x, y) \geq d \; \forall x, y \in \mathcal{Y}, x \neq y\} \; . \tag{1.9}$$

This function is familiar in coding theory; it gives the *maximal cardinality of a q-ary code of length n and minimum distance d*. The two complementary sets of binary vectors of length n with an even and odd number of ones, Z_+ and Z_-, are binary codes with minimum distance two and have cardinality $|Z_\pm| = \mathcal{A}_2(n, 2) = 2^{n-1}$. They are *perfect binary codes* of length n and minimum distance 2.

Theorem 1.3.1. (Mixtures of discrete product distributions). *Let* $\mathcal{X} = \bigtimes_{i\in[n]} \mathcal{X}_i$, $|\mathcal{X}_1| = \max\{|\mathcal{X}_i|\}$ *and* $|\mathcal{X}_n| = \min\{|\mathcal{X}_i|\}$.

- *If* $m \geq |\mathcal{X}|/|\mathcal{X}_1|$, *then* $\mathrm{Mixt}^m(\mathcal{E}^1) = \mathcal{P}$.
- *If* $\mathrm{Mixt}^m(\overline{\mathcal{E}^1}) \supseteq \mathcal{P}$, *then* $m \geq \mathcal{A}_q(n,2)$, *where* $q = |\mathcal{X}_n|$. *Furthermore,* $\mathcal{A}_q(n,2) \geq \frac{q^n}{1+n(q-1)}$ *and* $\mathcal{A}_q(n,2) = q^{n-1}$ *if* q *is a prime power.*

In particular, if q *is a prime power and* $\mathcal{X} = \{1,\ldots,q\}^n$, *then*

$$\mathrm{Mixt}^m(\mathcal{E}^1_{n,q\text{-ary}}) = \mathcal{P} \quad \textit{if and only if} \quad m \geq q^{n-1}\ .$$

Proof. We use Lemma 1.2.2. The following $|\mathcal{X}|/|\mathcal{X}_1|$ S-sets cover $\mathcal{X}$: $\{\{(x,y)\}_{x\in\mathcal{X}_1} : y \in \bigtimes_{i\in[n]\setminus\{1\}} \mathcal{X}_i\}$. For the second item: An edge of $\mathrm{cs}(\mathcal{E}^1)$ is given by a pair $\{A_x, A_y\}$ with q-ary vectors x and y of length n which differ in exactly one entry, $d_H(x,y) = 1$. A set $\mathcal{Y} \subset \mathcal{X}$ containing no such pair can be packed only using S-sets of cardinality one, because any facial set of cardinality larger than one always contains edges. If the minimum distance of a code is two, then obviously the code doesn't contain any edges. The *Gilbert-Varshamov bound* [49, 116] is: $\mathcal{A}_q(n,2) \geq \frac{q^n}{\sum_{j=0}^{d-1}\binom{n}{j}(q-1)^j}$, and in the prime power case it reads: $\mathcal{A}_q(n,d) \geq q^k$, where k is the largest integer with $q^k < \frac{q^n}{\sum_{j=0}^{d-2}\binom{n-1}{j}(q-1)^j}$. On the other hand, we have the *singleton bound* [106]: $\mathcal{A}_q(n,d) \leq q^{n-d+1}$. For $d = 2$ the combination of the two bounds completes the proof. □

Remark 1.3.2.

(i) Any distribution in $\overline{\mathcal{P}}$ can be approximated arbitrarily well by a mixture of q^{n-1} elements from $\mathcal{E}^1$, in view of Theorem 1.3.1 and $\overline{\mathrm{Mixt}^m(\mathcal{E})} = \mathrm{Mixt}^m(\overline{\mathcal{E}})$. Furthermore, $\overline{\mathcal{P}} \setminus \mathrm{Mixt}^m(\mathcal{E}^1)$ has non-empty interior whenever $m < \mathcal{A}_q(n,2)$.

(ii) The convex support of the independence model is not two-neighborly. A decomposition of $\mathcal{X}$ based only on the neighborliness of $\mathrm{cs}(\mathcal{E}^1)$ would yield $|\mathcal{X}|$ mixture components, instead of $|\mathcal{X}|/\max_i|\mathcal{X}_i|$.

(iii) We see that the approximation of arbitrary distributions in $\overline{\mathcal{P}}$ to an arbitrary accuracy requires a very large number of product mixture components. The expected dimension of the model $\mathrm{Mixt}^m(\mathcal{E}^1)$ is $\min\{\sum_{i=1}^n(|\mathcal{X}_i|-1)m + (m-1), (\prod_{i\in[n]}|\mathcal{X}_i|)-1\}$. In the case of q-ary variables, q a prime power, the smallest mixture model that contains the full probability simplex has $n(q-1)q^{n-1}+(q^{n-1}-1)$ parameters, a number which surpasses $\dim(\mathcal{P}_{n,q\text{-ary}}) = q^n - 1$ by $n\frac{q-1}{q} + q^{n-1} - 1$. The exact dimension of $\mathrm{Mixt}^m(\mathcal{E}^1_{n,q\text{-ary}})$ is to date unknown for general q. This is an interesting and active research topic in algebraic geometry [1, 26]. In Remark 1.3.4 we comment on the binary case.

In the case of binary variables we have the following:

Corollary 1.3.3. (Mixtures of binary product distributions). *The mixture model* $\mathrm{Mixt}^m(\mathcal{E}^1_{n,\text{bin}})$ *doesn't contain any probability distribution with support on a binary code of minimum distance at least two and cardinality more than* m. *Furthermore,*

$$\mathrm{Mixt}^m(\mathcal{E}^1_{n,\text{bin}}) = \mathcal{P}_{n,\text{bin}} \quad \textit{if and only if} \quad m \geq 2^{n-1}\ .$$

Proof. This is a special case of Theorem 1.3.1. We give a binary version of the proof: We use Lemma 1.2.2 and the fact that the support sets of probability distributions in $\overline{\mathcal{E}^1_{n,\text{bin}}}$ are faces of the n-cube. Any packing of a binary code of minimum distance more than two by faces of the cube consists of single points. If the code has cardinality k, then k points are needed to pack it. Any $p \in \overline{\mathcal{P}_{n,\text{bin}}}$ with support contained in the union of m edges of the n-cube is contained in $\text{Mixt}^m(\overline{\mathcal{E}^1_{n,\text{bin}}})$. There are 2^{n-1} edges covering all vertices of the cube. Any representation of any $p \in \overline{\mathcal{P}}(Z_{\pm,n})$ as mixture of product distributions has at least $|\text{supp}(p)|$ components with support of cardinality one. □

In Appendix 1.B we provide refinements of this result, in the sense that we describe portions of the complement of $\text{Mixt}^m(\mathcal{E}^1_{n,\text{bin}})$.

Remark 1.3.4. (Dimension and identifiability).

(i) For $n \geq 2$ the graph of the n-cube has more than $2^{2^{n-2}}$ perfect matchings. A perfect matching is a set of pairwise disjoint edges of a graph covering all its vertices. Hence mixture decompositions into sums of probability distributions supported by pairs with Hamming distance one are highly non-unique (although the number of possible decompositions is in many cases finite). The distributions with support on a code of distance at least two have a unique representation as mixture of independent points in $\mathcal{E}^1_{n,\text{bin}}$. This should be compared to a result by Bocci and Chiantini [22], which shows that if $n > 5$, then any *generic* point of $\text{Mixt}^m(\mathcal{E}^1_{n,\text{bin}})$ is contained in only one m-secant of $\mathcal{E}^1_{n,\text{bin}}$ (the affine hull of m independent points in $\mathcal{E}^1_{n,\text{bin}}$), for all $m \leq \frac{2^{n-1}}{n}$.

(ii) The (non-negative) *outer-product rank* of a tensor is the smallest number of (non-negative) rank-one tensors that can represent it as their sum. A rank-one tensor is an n-way table that can be written as a product $p_1 \otimes \cdots \otimes p_n$. A consequence of recent work by Catalisano, Geramita and Gimigliano [26] is that the model $\text{Mixt}^m(\mathcal{E}^1_{n,\text{bin}})$ always has the expected dimension, except for the case $m = 3$ and $n = 4$, where the model has dimension one less than expected. If $\tilde{m}$ denotes the smallest m for $\text{Mixt}^m(\mathcal{E}^1_{n,\text{bin}}) = \mathcal{P}$, and $\hat{m}$ denotes the smallest m for which the mixture model has the same dimension as the probability simplex, then our result implies $\tilde{m} = 2^{\lfloor \log_2(n+1) \rfloor - 1}\hat{m}$. Let $\text{Sec}^m(\mathcal{E}^1_{n,\text{bin}})$ denote the collection of all affine combinations of m points in $\mathcal{E}^1_{n,\text{bin}}$. The number $\hat{m}$ is the smallest m for which $\text{Sec}^m(\mathcal{E}^1_{n,\text{bin}})$ contains $\mathcal{P}$ (more precisely, the closure $\overline{\text{Sec}^m(\mathcal{E}^1_{n,\text{bin}})}$ in the Zariski topology contains $\mathcal{P}$), and equals (generically) the rank of $(2 \times \cdots \times 2)$-tensors. The number $\tilde{m}$ is the maximal non-negative outer-product rank of non-negative $(2 \times \cdots \times 2)$-tensors. The discrepancy between the two types of rank is called a *rank jump*.

The following is a consequence of Corollary 1.3.3:

Corollary 1.3.5. *Let* $1 \leq j \leq n-1$. *If* $\text{Mixt}^m(\overline{\mathcal{E}^1_{n,\text{bin}}}) \supseteq \mathcal{E}^j_{n,\text{bin}}$, *then*

$$m \geq \max\left\{\kappa^f_{1,j}, \left\lceil \frac{\dim(\mathcal{E}^j_{n,\text{bin}}) + 1}{n+1} \right\rceil \right\}.$$

Furthermore, $\kappa^f_{1,j} \geq \max\{|\mathcal{Z}| : \mathcal{Z} \in \mathcal{F}(\mathcal{E}^j_{n,\text{bin}}),\ \mathcal{Z} \subseteq Z_\pm\} \geq 2^j - 1$.

Proof. If $\mathrm{Mixt}^m(\overline{\mathcal{E}^1_{n,\mathrm{bin}}}) \supseteq \mathcal{E}^j_{n,\mathrm{bin}}$ it is necessary that $\dim(\mathrm{Mixt}^m(\mathcal{E}^1_{n,\mathrm{bin}})) \geq \dim \mathcal{E}^j_{n,\mathrm{bin}}$. Parameter counting yields $mn + m - 1 \geq \dim(\mathrm{Mixt}^m(\mathcal{E}^1_{n,\mathrm{bin}})) \geq \dim \mathcal{E}^k_{n,\mathrm{bin}}$. The quantity $\max\{|\mathcal{Z}| : \mathcal{Z} \in \mathcal{F}(\mathcal{E}^j_{n,\mathrm{bin}}) \text{ and } \mathcal{Z} \subseteq Z_{\pm}\}$ is lower bounded by $2^j - 1$, which can be seen from [64, Theorem 13]. □

Example 1.3.6.

(i) By Corollary 1.3.5 and Example 1.2.9: If $\mathrm{Mixt}^m(\overline{\mathcal{E}^1_{4,\mathrm{bin}}}) \supseteq \mathcal{E}^2_{4,\mathrm{bin}}$, then $m \geq 6$.

(ii) The polytope $\mathrm{cs}(\mathcal{E}^3_{4,\mathrm{bin}})$ has dimension 14 and $\left\lceil (\dim \mathcal{E}^3_{4,\mathrm{bin}} + 1)/(4+1) \right\rceil = 3$. On the other hand $2^3 - 1 = 7$. By Corollary 1.3.5 if $\mathrm{Mixt}^m(\overline{\mathcal{E}^1_{4,\mathrm{bin}}}) \supseteq \mathcal{E}^3_{4,\mathrm{bin}}$, then $m \geq 7$.

In later chapters we will derive analogous results for the inclusion of RBMs in mixtures of independence models.

Interaction Models

Now we turn our attention to mixtures of more general hierarchical models than independence models.

Remark 1.3.7. As explained in Section 1.1, each element from the hierarchical model $\mathcal{E}_\Delta(\mathcal{X}_1 \times \cdots \times \mathcal{X}_n)$ with interaction sets Δ factorizes according to $p \propto \prod_{\lambda\in\Delta} \exp(\phi_\lambda(x_\lambda))$. The model $\mathrm{Mixt}^m(\mathcal{E}_\Delta)$ can be understood as the set of marginal visible distributions from a hierarchical model $\mathcal{E}_{\widetilde{\Delta}}$ with $(n+1)$ variables; the n visible variables X_i, $i = 1, \ldots, n$, and one hidden variable X_{n+1}. The interaction sets of $\mathcal{E}_{\widetilde{\Delta}}$ are $\widetilde{\Delta} := \{\lambda \cup \{n+1\} \colon \lambda \in \Delta\} \cup \Delta \cup \{(n+1)\}$. The joint probability distributions have the following from: $p(x_1, \ldots, x_n, x_{n+1}) \propto \prod_{\lambda\in\Delta} \exp(\phi_{\lambda,(n+1)}(x_\lambda, x_{n+1}))$. The visible marginal distributions are of the form:

$$p(x_1, \ldots, x_n) = \sum_{x_{n+1}} \prod_{\lambda\in\Delta} \exp(\phi_\lambda^{x_{n+1}}(x_\lambda)) \exp(\phi(x_{n+1})) \,, \tag{1.10}$$

and can be written as $p(x_1, \ldots, x_n) = \sum_h \alpha(h) p^h(x_1, \ldots, x_n)$, where p^h is an arbitrary element of $\mathcal{E}_\Delta$ for each value of h, and α are arbitrary mixture weights. Hence the mixture model of a hierarchical model is the marginal of another hierarchical model. Figure 1.3 shows a factor-graph representation of the model $\mathcal{E}_{\widetilde{\Delta_2}}$ for four visible variables. See [118] for details on factor graphs.

Proposition 1.3.8. *Let $\mathcal{X} = \bigtimes_{i\in[n]} \mathcal{X}_i$. Let $\tilde{\kappa}_k(\mathcal{Y})$ denote the minimal cardinality of a covering of $\mathcal{Y} \subseteq \mathcal{X}$ using cylinder sets of dimension k. Consider any $p \in \overline{\mathcal{P}}$ and let $\kappa := \min\{\frac{|\mathrm{supp}(p)|}{2^k-1}, \tilde{\kappa}_k(\mathrm{supp}(p))\}$. If $m \geq \kappa$, then $p \in \mathrm{Mixt}^m(\overline{\mathcal{E}^k})$.*

Proof. This follows from Example 1.2.3, the $(2^k - 1)$-neighborliness of $\mathcal{E}^k$ (see [64]), and Lemma 1.2.2. □

The following is a stronger result for binary variables and our main result on mixtures of binary hierarchical models:

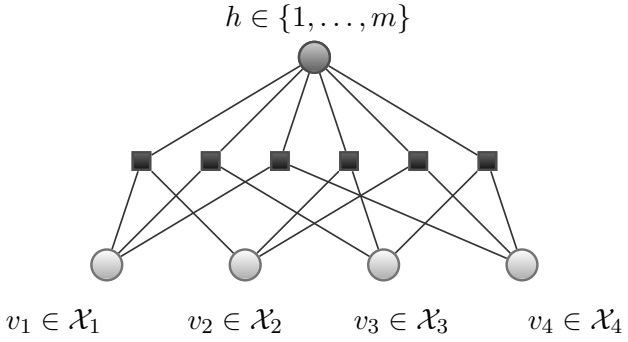

Figure 1.3: *Factor graph representation of the m-mixture model of the set of probability distributions of 4 variables $\{v_i\}_{i=1}^n$ taking values on $\{\mathcal{X}_i\}_{i=1}^4$ and involving pairwise interactions. The (circular) nodes represent variables and the black squares represent factors. The dark node at the top represents a hidden variable. The nodes connected to a common factor are fully interacting. If $\mathcal{X}_i = \{0, 1\}$ for all i, the depicted model is $\mathrm{Mixt}^m(\mathcal{E}^2_{4,\mathrm{bin}})$, and by Theorem 1.3.9, $m = 3$ is a sufficient number of states of h for which the model can represent any visible distribution.*

Theorem 1.3.9. (Mixtures of binary k-interaction models). *There is a covering of all vertices of $\mathrm{cs}(\mathcal{E}^k_{n,\mathrm{bin}})$ by $2^{n-(k+1)}(1 + \frac{1}{2^k-1})$ simplex faces. Hence:*

$$\textit{If } \; m \geq 2^{n-(k+1)}\big(1 + \frac{1}{2^k - 1}\big), \; \textit{then } \; \mathrm{Mixt}^m(\mathcal{E}^k_{n,\mathrm{bin}}) = \mathcal{P}_{n,\mathrm{bin}}.$$

In particular, if $p \in \overline{\mathcal{P}_{n,\mathrm{bin}}}$ and $m \geq \tilde{\kappa}_{k+1}(\mathrm{supp}(p))(1 + \frac{2}{2^k-1})$, then $p \in \mathrm{Mixt}^m(\overline{\mathcal{E}^k_{n,\mathrm{bin}}})$.

Remark 1.3.10.

(i) Theorem 1.3.9 holds for any $\mathcal{E}_\Delta(\{0, 1\}^n)$ with $\Delta \supseteq \Delta_k$.

(ii) The result halves the bound on m computed in Proposition 1.3.8 in the case of full support binary distributions.

(iii) For $k = 1$, the bound $2^{n-(k+1)}(1 + \frac{1}{2^k-1}) = 2^{n-1}$ recovers the upper bound on m given in Corollary 1.3.3.

(iv) For $n = 4$ and $k = 2$ the cardinality bound for a covering of the vertices of $\mathrm{cs}(\mathcal{E}^k_{n,\mathrm{bin}})$ by simplex faces given in Theorem 1.3.9 is $\lceil 2^{4-(2+1)}/(1 - 2^{-2}) \rceil = 3$. This is optimal, in view of Example 1.2.9.

Before proving Theorem 1.3.9 we need to elaborate the components of the proof. We will use the S-sets of cardinality $2(2^k - 1)$ described in Lemma 1.3.12.

Kahle [64] investigates the neighborliness of marginal polytopes and shows that the convex support of $\mathcal{E}^k$ is $(2^k - 1)$-neighborly. For binary variables this result gives the maximal neighborliness degree of $\mathrm{cs}(\mathcal{E}^k_{n,\mathrm{bin}})$, because there exist sets of cardinality 2^k which are not S-sets of $\mathcal{E}^k_{n,\mathrm{bin}}$:

Proposition 1.3.11. *Let $\mathcal{X} = \{0, 1\}^n$ and $0 < k < n$. Consider $\mathcal{E}^k_{n,\mathrm{bin}}$ and any $y_{\lambda^c} \in \mathcal{X}_{\lambda^c}$, $\lambda \subseteq [n], |\lambda| = k + 1$. Then $|[y_{\lambda^c}] \cap Z_\pm| = 2^k$, and any $\mathcal{Y} \subseteq \mathcal{X}$ containing $[y_{\lambda^c}] \cap Z_\pm$ is not an S-set. If $\mathcal{Y} \supseteq [y_{\lambda^c}] \cap Z_\pm$ and $\mathcal{Y} \not\supseteq [y_{\lambda^c}]$, then $\mathcal{Y}$ is not facial.*

Proof. See Appendix 1.A. □

On the other hand, if Q is a K-neighborly d-dimensional polytope with $d \geq 2K$, then every face F of Q with $0 \leq \dim F < 2K$ is a simplex, i.e., Q is $(2K-1)$-simplicial, see [51, Theorem 7.4.3]. This implies that all $(2K-1)$-dimensional faces of $\mathrm{cs}(\mathcal{E}^k)$ are simplices, where $K = 2^k - 1$. If $K > \left\lfloor \frac{1}{2} \dim(\mathrm{cs}(\mathcal{E}^k)) \right\rfloor$, then $\mathrm{cs}(\mathcal{E}^k)$ is a simplex (this only occurs when $k = n$ and $\overline{\mathcal{E}^n} = \overline{\mathcal{P}}$).

In the following we focus on the binary case and search for facial sets of $\mathcal{E}^k_{n,\mathrm{bin}}$ of cardinality $2K$, $K = 2^k - 1$ (which are S-sets). For $0 < k < n$, any $(k+1)$-dimensional cylinder set $[y_{\lambda^c}], \lambda \subseteq [n], |\lambda| = k+1$ is facial for $\mathcal{E}^k_{n,\mathrm{bin}}$ (it is a facial set of $\mathcal{E}^1_{n,\mathrm{bin}}$) and hence the vertices of $\mathrm{cs}(\mathcal{E}^k_{n,\mathrm{bin}})$ can be covered by $2^{n-(k+1)}$ disjoint faces $\{F_i\}_i$. $2K$ of the $(2K+2)$ vertices of each F_i are covered by a simplex face. Any polytope which is not a simplex always contains two disjoint faces of complementary dimension [34], such that the two additional vertices can be chosen as an edge of F_i. These two vertices in each F_i can be covered using the fact that $\mathrm{cs}(\mathcal{E}^k_{n,\mathrm{bin}})$ is K-neighborly, but also arranging the simplex faces conveniently. To this end use the following lemma, which collects similar results from [51, 61, 66]. We provide a thorough proof in Appendix 1.A.

Lemma 1.3.12. *Let $0 < k < n$ and $\mathcal{X} = \{0,1\}^n$. Any $(k+1)$-dimensional cylinder set $\mathcal{Y} \subseteq \mathcal{X}$ is a facial set of $\mathcal{E}^k_{n,\mathrm{bin}}$ and the corresponding face F of the convex support is a simplicial polytope combinatorially equivalent to the cyclic polytope $C(2^{k+1}, 2^{k+1}-2)$. There are exactly 2^{2k} S-sets of cardinality $(2^{k+1}-2)$ contained in $\mathcal{Y}$. The S-sets contained in $\mathcal{Y}$ are $\{\mathcal{Z} \subset \mathcal{Y} : Z_\pm \not\subseteq \mathcal{Z}\}$.*

Now we are ready for proving Theorem 1.3.9:

Proof of Theorem 1.3.9. Consider the following partition of $\{0,1\}^n$ into $(k+1)$-dim cylinder sets:

$$\{C_y\}_y := \{(x_1^{k+1}, x_{k+2}^n) \in \{0,1\}^n \colon x_{k+2}^n = y\}_{y \in \{0,1\}^{n-(k+1)}}$$

By Lemma 1.3.12 for any $y \in \{0,1\}^{n-(k+1)}$ the elements of C_y are disjointly covered by:
(i) An S-set of $\overline{\mathcal{E}^k}$ of cardinality $2K$. We denote this set by G_y.
(ii) A pair E_y, which can be chosen to be any edge of C_y (a pair differing in one entry), in particular:

$$E_y = \{(z_1^k, x_{k+1}, y) \in \{0,1\}^n \colon z_1^k \text{ fixed } \} \,. \tag{1.11}$$

The vector z can be chosen to be the same for all E_y, such that the S-sets $\{G_y\}_y$ satisfy:

$$\bigcup_{y \in \{0,1\}^{n-(k+1)}} G_y \quad = \quad \{0,1\}^n \setminus \tilde{C}_{n-k} \,,$$

where $\tilde{C}_{n-k}$ is the following $(n-k)$-dimensional cylinder set:

$$\tilde{C}_{n-k} = \bigcup_{y \in \{0,1\}^{n-(k+1)}} E_y \quad = \quad \{(z_1^k, \tilde{y}_1^{n-k}) : z_1^k \text{ fixed}\} \,.$$

The set $\tilde{C}_{n-k}$ can be considered as new state space which still has to be covered using S-sets. If $n-k < k+1$, only one S-set is required. Iteration until exhausting all coordinates yields that κ,

the minimal number of faces of $\mathrm{cs}(\mathcal{E}^k_{n,\mathrm{bin}})$ which are simplices and suffice to cover all vertices, is not more than:

$$\kappa \leq 1 + \sum_{0 \leq i \leq \frac{n-(k+1)}{k}} \frac{2^{n-ik}}{2^{k+1}} = \left\lceil \frac{2^n}{2^{k+1}} \sum_{i=0}^{\infty} \frac{1}{(2^k)^i} \right\rceil = \left\lceil \frac{2^{n-(k+1)}}{1-2^{-k}} \right\rceil . \qquad \square$$

We conclude this section with a few comments on the S-sets of binary hierarchical models:

From Proposition 1.3.11 we can derive an upper bound for the cardinality of an S-set. The *covering radius* $R(\mathcal{Y})$ of a binary code $\mathcal{Y} \subseteq \{0,1\}^n$ denotes the maximal Hamming distance from some $x \in \{0,1\}^n$ to $\mathcal{Y}$, i.e., $\max_{x\in\{0,1\}^n} \min_{y\in\mathcal{Y}} d_H(x,y)$. Let $K(n,k+1)$ denote the smallest cardinality of a binary code with length n and covering radius $k+1$. Let $B_{n,k+1}$ denote a Hamming ball of radius $k+1$ in $\{0,1\}^n$. We can give a (coarse) bound on the cardinality of S-sets as follows:

Proposition 1.3.13. *If $\mathcal{Y} \subseteq \mathcal{X}$ is an S-set of $\mathcal{E}^k$, then $|\mathcal{Y} \cap Z_\pm| \leq 2^{n-1} - K(n,k+1) \leq 2^{n-1}(1 - 2/|B_{n,k+1}|)$, and $|\mathcal{Y}| \leq |\Delta_k|$. Hence also $|\mathcal{Y}| \leq 2^n - 2K(n,k+1) \leq 2^n(1 - 2/|B_{n,k+1}|)$.*

Proof. See Appendix 1.A. $\square$

Example 1.3.14. For $n = 4$ and $k = 2$ Proposition 1.3.13 implies that any S-set $\mathcal{Y}$ satisfies $|\mathcal{Y} \cap Z_\pm| \leq 8 - 2 = 6$. In view of the computations from Example 1.2.9, the bound of Proposition 1.3.13 is sharp in this case.

In general, the estimation of upper bounds for $K(n,k)$ is a hard problem, see [77, 91, 27]. This complicates the specification of the optimal (necessary) number of S-sets which cover $\mathcal{X}$.

1.A Proofs and Details

Proof of Proposition 1.2.8. Assume w.l.o.g. that the sufficient statistics contains the row $\mathbb{1}$. The image of the map $\pi\colon p \mapsto A \cdot p$ restricted to $\overline{\mathcal{P}}$ is the convex support $Q := \mathrm{conv}\{A_x\}_{x\in\mathcal{X}}$. We denote by p_η the unique preimage of $\eta \in Q$ by the restricted moment map, $p_\eta = (\pi|_{\overline{\mathcal{E}}})^{-1}(\eta)$. The 2-mixture of $\overline{\mathcal{E}}$ is parametrized by a *mixture map* in the following way $\phi\colon D := \overline{\mathcal{P}}_2 \times Q^2 \to \overline{\mathcal{P}}\ ;\ (\alpha,\eta_1,\eta_2) \mapsto \sum_{i=1}^2 \alpha_i p_{\eta_i}$. From $\mathrm{Mixt}^2(\overline{\mathcal{E}}) \supset \partial\mathcal{P} := \overline{\mathcal{P}} \setminus \mathring{\mathcal{P}}$ it follows that the restriction $\phi|_C\colon C := \partial(\overline{\mathcal{P}}_2 \times Q^2) \to \partial\mathcal{P}$ is a continuous surjection. Consider the *normal space* of $\mathcal{E}$, which is given by $\mathcal{N} = \ker A$. For any $p \in \overline{\mathcal{P}}$ the linear model $\mathcal{N}_p := \{q \in \overline{\mathcal{P}} : p - q \in \mathcal{N}\}$ intersects $\overline{\mathcal{E}}$ at a unique point $p_{\mathcal{E}} \in \overline{\mathcal{E}} \cap \mathcal{N}_p$ (see [95, Theorem 2.16]). Hence, $\overline{\mathcal{P}} = (\overline{\mathcal{E}} + \ker A) \cap \overline{\mathcal{P}} = \cup_{p\in\overline{\mathcal{E}}} \mathcal{N}_p$. For every $p \in \mathcal{P}$, $\mathcal{N}_p$ is a polytope of dimension $\dim\ker A$. In the present case $\dim\ker A = 2$. The boundary of $\mathcal{N}_p$ is contained in the boundary of $\mathcal{P}$. Now, for any $p \in \mathcal{E}_\pentagon$ we consider the subset $B_p = \phi^{-1}(\mathcal{N}_p) = \{(\alpha,\eta_1,\eta_2) \in D\colon \sum_{i=1}^2 \alpha_i\eta_i = \pi(p)\}$. This set is mapped by ϕ to all convex combinations of 2 elements of $\overline{\mathcal{E}}_\pentagon$ which have the same expectation parameter as p. We consider also $\partial B_p = B_p \cap (\overline{\mathcal{P}}_2 \times (\partial Q)^2)$, which corresponds to the same kind of mixtures, but with mixture components from the boundary of $\mathcal{E}$. We have that $\phi\colon \partial B_p \to \partial\mathcal{N}_p$ is surjective and has degree 2! (this is the cardinality of the preimage of a regular value, which arises from the freedom to permute the mixture components). For $Q_\pentagon$ we see that ∂B_p is parametrized by an angle, say γ, and that $\phi|_{\partial B_p}(\gamma)$ circulates $\partial\mathcal{N}_p$ twice. Using that B_p is contractible it follows that $\phi|_{B_p} = \mathcal{N}_p$ and $\mathrm{Mixt}^2(\overline{\mathcal{E}}) = \overline{\mathcal{P}}$. For strictly positive distributions the claim follows from

the fact that $\phi(\mathring{B}_p) \subseteq \mathcal{P}$ and that the image of an ε-retraction of B_p, $(1-\varepsilon)(B_p - p) + p$, can be made such that it contains any δ-retraction of $\mathcal{N}_p$, $(1-\delta)(\mathcal{N}_p - p) + p$. □

Topological Discussion. We abbreviate $\mathrm{cs}(\mathcal{E})$ by Q. Consider the following *mixture map*

$$\tilde{\phi} \;:\quad \overline{\mathcal{P}}_m \times Q^m \to \overline{\mathcal{P}}\;; \quad (q, \eta_1, \ldots, \eta_m) \mapsto \sum_{i=1}^{m} q_i p_{\eta_i}\;,$$

where $(q_i)_{i\in[m]}$ are the mixture weights and $p_\eta \in \overline{\mathcal{E}}$ is the unique preimage of $\eta \in Q$ by the moment map $\overline{\mathcal{E}} \xrightarrow{\sim} Q \;;\; p \mapsto A \cdot p$. The parameters η are also known as *expectation parameters* [10]. We can also consider the restriction ϕ of $\tilde{\phi}$ to a domain which is homeomorphic to the *join* of Q with itself $Q^{\star m}$, (the join $U \star V$ of U and V is the quotient space $A \times B \times [0,1]/\sim$, where $a \times b \times 0 \sim a \times \tilde{b} \times 0$ for all $a \in A$ and $b, \tilde{b} \in B$ and $a \times b \times 1 \sim \tilde{a} \times b \times 1$ for all $a, \tilde{a} \in A$ and $b \in B$). This identifies all values of an entry η_i when $q_i = 0$.

If the mixture model contains $\partial\mathcal{P}$, then the idea of using topology is to show that a retraction of $\phi^{-1}(\partial\mathcal{P})$ within $Q^{\star m}$ in fact induces by ϕ a retraction of $\partial\mathcal{P}$ on the image of ϕ which doesn't leave out any points of $\mathcal{P}$. This is formalized in the following Proposition 1.A.1. Here we denote by $[\gamma]$ the homotopy class of γ, by S^n the n-dimensional sphere, and by $\pi_n(X)$ the n-th homotopy group of X.

Proposition 1.A.1. *Consider a $(n+1)$-dim simplex $\overline{\mathcal{P}}$, a contractible set D and a continuous map $\phi\colon D \to \overline{\mathcal{P}}$. If there exists a map $\gamma : S^n \to D$ such that $[\phi \circ \gamma] \in \pi_n(\partial\mathcal{P}) \setminus 0$, e.g., if there exists a $C \subseteq D$ s.t. $\phi|_C\colon C \to \partial\mathcal{P}$ is a covering space projection (a covering map), then $\phi(D) = \overline{\mathcal{P}}$.*

Proof. The existence of the map γ means that there is a subset $C \subseteq D$ such that $\phi|_C\colon C \to \partial\mathcal{P} \simeq S^n$ and the induced homomorphism of the n-th homotopy group $(\phi|_C)_*\colon \pi_n(C) \to \pi_n(\partial\mathcal{P})$ satisfies $\mathrm{Im}(\phi|_C)_* \supsetneq 0$. The condition that the image of $(\phi|_C)_*$ is not the trivial group implies that $\phi|_C\colon C \to \partial\mathcal{P}$ is surjective (any map $S^n \to S^n$ which is not surjective is null-homotopic). If we assume that $\phi(D) \neq \overline{\mathcal{P}}$, then there exists a $y \in \mathcal{P} \setminus \partial\mathcal{P}$ with $y \notin \phi(D)$. But in this case $\pi_n(\phi(D)) \neq 0$ and in particular any $g : S^n \to \partial\mathcal{P}$ which is not null-homotopic in $\partial\mathcal{P}$ isn't null-homotopic in $\phi(D)$ either. This is a contradiction: Since $\mathrm{Im}(\phi|_C)_*$ contains one non-trivial homotopy class, there is one $[g] \in \pi_n(\partial\mathcal{P}) \setminus \{0\}$ for which $(\phi|_C)_*^{-1}([g])$ exists. On the other hand, any element of $[\tilde{g}] \in (\phi|_C)_*^{-1}([g])$ is null-homotopic in D, because D is contractible. Via ϕ this yields the null-homotopy of $[g]$. For the example: If $\phi|_C$ is a covering space projection onto $\partial\mathcal{P}$, then the induced maps of homotopy-groups $(\phi|_C)_* : \pi_n(C) \to \pi_n(\partial\mathcal{P})$ are isomorphisms for all $n \geq 2$, [53, Proposition 4.1]. □

For completeness we provide here a more extensive characterization of S-sets. For any $m \in \mathbb{R}^{\mathcal{X}}$ let $m^{\pm}(x) := \max\{0, \pm m(x)\}$ and let $p^m := \prod_{x\in\mathcal{X}} (p(x))^{m(x)}$. We use the following results:

Theorem 1.A.2. (Geiger et al. [48] and Rauh et al. [96]). *Let $\mathcal{E}$ be an exponential family with sufficient statistics $A \in \mathbb{R}^{d\times\mathcal{X}}$. (I) A set $\mathcal{Y} \subseteq \mathcal{X}$ is facial iff there exists one $p \in \overline{\mathcal{E}}$ with $\mathrm{supp}(p) = \mathcal{Y}$ iff $\mathrm{supp}(m^+) \subseteq \mathcal{Y} \Leftrightarrow \mathrm{supp}(m^-) \subseteq \mathcal{Y}$ for all $m \in \ker A$. (II) A distribution p is contained in $\overline{\mathcal{E}}$ iff p fulfills the equations $p^{m^+} - p^{m^-} = 0 \;\; \forall m \in \ker A$.*

For simplicity we focus on the binary case. Consider an exponential family $\mathcal{E}$ on $\mathcal{X}$ with sufficient statistics $A(\Delta, \mathcal{X}) = (A(\lambda, x))_{\lambda\in\Delta, x\in\mathcal{X}}$, which includes the row $A(\emptyset, \mathcal{X}) = \mathbb{1}$. Let

$A(\Delta^c, \mathcal{X})$ span the orthogonal complement of $A(\Delta, \mathcal{X})$.

Proposition 1.A.3. *Given any $\mathcal{Y} \subseteq \mathcal{X}$ the following statements are equivalent:*

(i) *Every $p \in \overline{\mathcal{P}}$ with $\operatorname{supp}(p) = \mathcal{Y}$ is contained in $\overline{\mathcal{E}}$.*

(ii) *Every $p \in \overline{\mathcal{P}}$ with $\operatorname{supp}(p) \subseteq \mathcal{Y}$ is contained in $\overline{\mathcal{E}}$.*

(iii) $\operatorname{supp}(m^+) \not\subset \mathcal{Y} \;\; \forall m \in \ker A(\Delta, \mathcal{X}) \setminus \{0\}$.

(iv) $\operatorname{rk} A(\Delta^c, \mathcal{Y}^c) = |\Delta^c|$ *and $\mathcal{Y}$ is facial.*

(v) $\operatorname{rk} A(\Delta, \mathcal{Y}) = |\mathcal{Y}|$ *and $\mathcal{Y}$ is facial.*

(vi) *Every $\mathcal{Y}' \subseteq \mathcal{Y}$ is facial. I.e., $\mathcal{Y}$ corresponds to a $(|\mathcal{Y}| - 1)$-simplex face of the convex support; $\{A(\Delta, y)\}_{y \in \mathcal{Y}}$ are the vertices of a $(|\mathcal{Y}| - 1)$-simplex face of $\operatorname{conv}\{A_x\}_{x \in \mathcal{X}}$.*

Proof. The equivalence of *(i)* and *(ii)* is trivial. The claim *(ii) if (iii)* follows directly from Theorem 1.A.2 (II). For the implication *only if* we have to show that if $\operatorname{supp}(m^+) \cap \mathcal{Y}^c = \emptyset$ for some $m \in \ker A(\Delta, \mathcal{X}) \setminus \{0\}$, then there exists a $p \in \overline{\mathcal{P}}$ with $\operatorname{supp}(p) = \mathcal{Y}$ and $p^{n^+} - p^{n^-} \neq 0$ for some $n \in \ker A(\Delta, \mathcal{X})$. Assume $\operatorname{supp}(m^+) \subseteq \mathcal{Y}$. If there exists one $\tilde{p} \in \overline{\mathcal{E}}$ with support $\mathcal{Y}$ (if none exists we are done), then $\mathcal{Y}$ is facial and $\operatorname{supp}(m^-) \subseteq \mathcal{Y}$. We write $(\tilde{p}_i)_{\{i : m_i \neq 0\}} = (\tilde{\xi}, \tilde{\eta}) \in \mathbb{R}_+^{\operatorname{supp}(m^+)} \times \mathbb{R}_+^{\operatorname{supp}(m^-)}$. $|\operatorname{supp}(m^\pm)| > 0$, since $0 = \langle A(\emptyset, \mathcal{X}), m \rangle = \sum_x m(x)$. Assume $\|\tilde{\xi}\|_1 < \|\tilde{\eta}\|_1$, (if this is not possible, again we are done). By (II) $\tilde{\xi}^{m^+} - \tilde{\eta}^{m^-} = 0$. Now consider a p with $p(x) = \tilde{p}(x)\ \forall x : m(x) = 0$, and $\xi = 2\tilde{\xi}$, and $\eta = (1 - \|\tilde{\xi}\|_1 / \|\tilde{\eta}\|_1)\tilde{\eta}$ in the other entries. We have $\|\xi\|_1 + \|\eta\|_1 = \|\tilde{\xi}\|_1 + \|\tilde{\eta}\|_1$ s.t. $p \in \overline{\mathcal{P}}$, and:

$$\xi^{m^+} - \eta^{m^-} = \left(2^{\langle \mathbb{1}, m^+ \rangle} - \left(1 - \|\tilde{\xi}\|_1 / \|\tilde{\eta}\|_1\right)^{\langle \mathbb{1}, m^- \rangle}\right) \tilde{\xi}^{m^+}.$$

W.l.o.g. $\langle \mathbb{1}, m^\pm \rangle > 1$. Since $0 < \|\tilde{\xi}\|_1 / \|\tilde{\eta}\|_1 < 1$ and $\tilde{\xi}$ is greater than 0 in every entry, $\xi^{m^+} - \eta^{m^-} \neq 0$.

(iv) iff *(iii)*: It suffices to show: For a facial $\mathcal{Y}$ it is $\operatorname{rk} A(\Delta^c, \mathcal{Y}^c) = |\Delta^c|$ if and only if $\mathcal{Y}^c \cap \operatorname{supp}(m) \neq \emptyset \;\; \forall m \in \ker A(\Delta, \mathcal{X}) \setminus \{0\}$. Any $m \in \ker A(\Delta, \mathcal{X})$ can be written as $m(x) = \langle \alpha, A(\Gamma, x) \rangle$, where $\operatorname{supp}(\alpha) \subseteq \Delta^c$. For any $x \in \mathcal{X}$ $m(x) = \langle \alpha, A(\Gamma, x) \rangle = 0 \;\Leftrightarrow\; \alpha \perp A(\Gamma, x)$. Hence, $\mathcal{Y}^c \cap \operatorname{supp}(m) = \emptyset$ is equivalent to the existence of some $\alpha \in \mathbb{R}^\Gamma$ such that $\alpha \perp A(\Gamma, x)\ \forall x \in \mathcal{Y}^c$. These equations can't be satisfied for any $\alpha \neq 0$ with $\operatorname{supp}(\alpha) \subseteq \Delta^c$ iff $\operatorname{rk} A(\Delta^c, \mathcal{Y}^c) = |\Delta^c|$.

(v) iff *(iv)*: We show $\operatorname{rk} A(\Delta, \mathcal{Y}) = \min\{|\mathcal{Y}|, |\Delta|\}$ iff $\operatorname{rk} A(\Delta^c, \mathcal{Y}^c) = \min\{|\mathcal{Y}^c|, |\Delta^c|\}$. Consider first $|\mathcal{Y}| = |\Delta|$. It suffices to show one direction, since one may define $\Delta' = \Delta^c, \mathcal{Y}' = \mathcal{Y}^c$. If $A(\Delta, \mathcal{Y})$ has full rank, then there exist two invertible $|\Delta| \times |\Delta|$-matrices L and R such that $LA(\Delta, \mathcal{Y})R = I_{|\Delta|}$. Now, multiplication of A with the block diagonal concatenation of L and R with $I_{|\Delta^c|}$ and appropriate row and column addition gives $\operatorname{diag}(I_{|\Delta|}, A(\Delta^c, \mathcal{Y}^c))$. The rank of this matrix is the same as that of A, and hence $\operatorname{rk} A(\Delta^c, \mathcal{Y}^c) = |\Delta^c|$. Consider now the case $|\mathcal{Y}| \neq |\Delta|$. W.l.o.g. $|\mathcal{Y}| \leq |\Delta|$ and $\operatorname{rk} A(\Delta, \mathcal{Y}) = |\mathcal{Y}|$. Since $A(\Delta, \mathcal{X})$ has full rank $|\Delta|$, there exists a set $\tilde{\mathcal{Y}}$ s.t. $\mathcal{X} \supseteq \tilde{\mathcal{Y}} \supseteq \mathcal{Y}$, $|\tilde{\mathcal{Y}}| = |\Delta|$ and $\operatorname{rk} A(\Delta, \tilde{\mathcal{Y}}) = |\Delta|$. From the first part we have that this is equivalent to $\operatorname{rk} A(\Delta^c, \tilde{\mathcal{Y}}^c) = |\Delta^c|$. But this implies $\operatorname{rk} A(\Delta^c, \mathcal{Y}^c) = |\Delta^c|$, since $\mathcal{Y}^c \supseteq \tilde{\mathcal{Y}}^c$. The other direction is analogue.

(vi) iff *(v)*: $\operatorname{rk} A(\Delta, \mathcal{Y}) = |\mathcal{Y}|$ is equivalent to $\{A(\Delta, y)\}_{y \in \mathcal{Y}}$ being linearly independent, such that $\operatorname{conv}\{A(\Delta, y)\}_{y \in \mathcal{Y}}$ is a $(|\mathcal{Y}| - 1)$-simplex. If $\mathcal{Y}$ is facial, then all sets $\mathcal{Y}' \subseteq \mathcal{Y}$ are facial. If

$\mathrm{conv}\{A(\Delta, y)\}_{y\in\mathcal{Y}}$ is a simplex face of Q, then $\{A(\Delta, y)\}_{y\in\mathcal{Y}}$ are affinely independent and in fact linearly independent. □

Proof of Proposition 1.3.11. From Lemma 1.2.5 we have: $\mathcal{Y}$ is not an S-set $\Leftrightarrow \exists m \in \ker A(\Delta, \mathcal{X})\setminus \{0\}$ such that $\mathrm{supp}(m^+) \subseteq \mathcal{Y}$. If $\exists m \in \ker A(\Delta, \mathcal{X}) \setminus \{0\}$ such that $\mathrm{supp}(m^+) = \mathcal{Y}$, then $\mathcal{Y}$ is not facial (from [48, 96]). From [64] we have that $\forall\Lambda \subseteq [n]$ with $|\Lambda| = k + 1$ and $\forall y \in \{0, 1\}^{n-(k+1)}$ there exists an $m \in \ker A(\Delta_k, \mathcal{X})$ such that $\mathrm{supp}(m) = \{x \in \mathcal{X} : x_{\Lambda^c} = y\} =: C$, (which is a $(k + 1)$-dim face of the n-cube), and $m|_C \propto A(\Lambda, C)$. Now observe that $A(\Lambda, x)$ is just the parity of x_Λ, i.e., $A(\Lambda, x)$ is 1, if x_Λ has an even number of ones and -1, if x_Λ has an odd number of ones. This means that $\mathrm{supp}(m^+) = Z_+ \cap C = \{x \in \mathcal{X} : \sum_{i\in\Lambda} x_i \mod 2 = 0, x_{\Lambda^c} = y\}$. Hence the claim. □

Proof of Lemma 1.3.12. The set $\mathcal{Y}$ has cardinality 2^{k+1} and therefore F has dimension at most $2^{k+1} - 1 = 2K + 1$. In fact it has a dimension strictly less than $2K + 1$, since in that case it would be a simplex, in contradiction to Proposition 1.3.11. On the other hand, if the dimension of F was less than $2K$, then by the arguments from page 26 (see [51, Theorem 7.4.3]), it would be a simplex, in contradiction to the number of vertices. Hence, $\dim F = 2^{k+1} - 2$, and all proper faces of F are simplices. The combinatorial equivalence of F to the cyclic polytope $C(2K + 2, 2K)$ follows from the fact that *Any $2n$-dimensional, n-neighborly polytope with $v \leq 2n + 3$ vertices is combinatorially equivalent to the cyclic polytope $C(v, 2n)$* (see [51, Theorem 7.2.3]). To complete the proof we use Gale's Evenness Criterion: *A d-tuple $V_J = \{x(t_j)\}_{j\in J}$ $J \subset [v], |J| = d$ of vertices of $C(v, d)$, spans a facet iff between any two elements of J there is an even number of elements in $[v] \setminus J$* (see [51, Theorem 4.7.2]). Here we have $v = 2^{k+1}$ and $d = 2^{k+1}-2$. The combinatorial structure of the cyclic polytope is independent of the map $i \mapsto t_i$ and we may choose $I = [v] := \{1, \ldots, 2^{k+1}\} \subset \mathbb{N}$. The sets $V_J, |J| = 2^{k+1} - 2$ satisfying the evenness criterion are exactly the complements of pairs $\{i^e, i^o\} \subset [v]$, where i^e is even and i^o is odd. There are 2^{2k} such pairs, and hence of facets. This is the same number of sets respecting the condition on S-sets from Proposition 1.3.11. Therefore, all sets $\mathcal{Z}$ respecting that condition, $\mathcal{Z} \not\supseteq \mathcal{Y} \cap Z_\pm$, must correspond to facets of $C(2^{k+1}, 2^{k+1} - 2)$ and are indeed S-sets. □

Proof of Proposition 1.3.13. For any S-set $\mathcal{Y}$ of $\mathcal{E}^k$ we have: $|(C\cap Z_\pm)\setminus\mathcal{Y}| \geq 1$ for any $(k+1)$-dimensional face of the n-cube C. Therefore, the maximal cardinality of an S-set $\mathcal{Y} \subseteq Z_\pm$ is upper bounded by $|Z_\pm| - K(n, k + 1)$, where $K(n, k + 1)$ is the smallest number of elements needed to mark all $(k + 1)$-dim faces of the n-cube. The set of vertices of all $(k + 1)$-faces of the n-cube containing a common mark x correspond to the Hamming ball $B_{n,k+1}(x) \subseteq \mathcal{X}$ of radius $k + 1$ centered at x. Hence, $K(n, k + 1)$ is the minimal cardinality of binary codes of length n and covering radius $k + 1$. In the case $R < n \leq 2R + 1$, clearly $K(n, R) = 2$, but in general computing $K(n, R)$ is hard (see [27]). A crude lower bound is the *sphere-covering bound*: $K(n, R) \geq 2^n/|B_{n,R}|$, which is only optimal if the faces containing different marks can be chosen to be disjoint. Here $|B_{n,R}| = \sum_{i=0}^{R} \binom{n}{i}$. On the other hand, the cardinality of an S-set of $\mathcal{E}^k$ can't exceed $\dim \mathrm{cs}(\mathcal{E}^k) + 1 = |\Delta_k| = |B_{n,k}|$, since the dimension of the corresponding face can't be larger than that of $\mathrm{cs}(\mathcal{E}^k)$. □

Details to Remark 1.2.10. The set $\mathcal{Y}$ is an S-set iff

(i) $\mathrm{rk}\, A(\Delta, \mathcal{Y}) = |\mathcal{Y}|$, (i.e., $\mathcal{Y}$ describes a $(|\mathcal{Y}| - 1)$-simplex), and

(ii) $\exists c \in \mathbb{R}^{|\Delta|}$ s.t. $\langle c, A(\Delta, y)\rangle = 0\ \forall y \in \mathcal{Y}$ and $\langle c, A(\Delta, x)\rangle \geq 1\ \forall x \in \mathcal{Y}^c$, (i.e., $\mathcal{Y}$ is facial).

We show that $\mathcal{Y}$ satisfies these properties iff $x * \mathcal{Y}$ does. We have that

$$A(\lambda, x * y) = (-1)^{|(\mathrm{supp}(x) \triangle \mathrm{supp}(y)) \cap \lambda|} \\ = (-1)^{|\mathrm{supp}(x) \cap \lambda|}(-1)^{|\mathrm{supp}(y) \cap \lambda|} \ , \ \ \forall x \in \mathcal{X}, \lambda \in 2^{[N]}, y \in \mathcal{X}$$

and thus $A(\Delta, x*\mathcal{Y}) = \mathrm{diag}\,(A(\Delta, x)) \cdot A(\Delta, \mathcal{Y})$. Hence, $\mathrm{rk}\, A(\Delta, \mathcal{Y}) = \mathrm{rk}\, A(\Delta, x*\mathcal{Y})$. On the other hand we can define $\tilde{c} := \mathrm{diag}(A(\Delta, x)) \cdot c$, and we get $\langle \tilde{c}, A(\Delta, x * y)\rangle = \langle c, A(\Delta, y)\rangle = 0\ \forall y \in \mathcal{Y}$. Similarly, $\langle \tilde{c}, A(\Delta, z')\rangle \geq 1\ \forall z' \in (y * \mathcal{Y})^c$. □

1.B Modes of Binary Mixture Models

The support sets of mixtures of distributions from the closure of an exponential family give information about the *topography* of the mixture model, i.e., about the kinds of *modes* that can be realized by the mixture model. In the following we elaborate this idea and estimate the volume of the complement of mixtures of binary independence models.

Definition 1.B.1. We call $x \in \{0,1\}^n$ a *mode* of $p \in \overline{\mathcal{P}_n}$ if $p(x) > p(\hat{x})$ for all $\hat{x}$ with $d_H(\hat{x}, x) = 1$, and we call x a *strong mode* if $p(x) > \sum_{\hat{x}: d_H(\hat{x},x)=1} p(\hat{x})$. We denote $\mathcal{G}_{n,m}$ the set of probability distributions in $\overline{\mathcal{P}_n}$ which have at least m modes, and $\mathcal{H}_{n,m}$ the set of probability distributions which have at least m strong modes.

Remark 1.B.2. Any strong mode is also a mode. Furthermore, a probability distribution on $\{0,1\}^n$ can have at most 2^{n-1} modes, because the Hamming distance between two modes is at least two. Hence $\overline{\mathcal{P}_n} = \mathcal{G}_{n,0} \supset \mathcal{G}_{n,1} \supset \cdots \supset \mathcal{G}_{n,2^{n-1}+1} = \emptyset$.

The modes of a probability distribution encode events that are locally most likely in the space of possible events. They are closely related to the possible support sets of probability distributions in statistical models, a problem that has been studied especially for hierarchical and graphical models without hidden variables [48, 66, 96]. The set of binary distributions with a fixed set of modes is a convex polytope inscribed in the probability simplex $\mathcal{P}_n$. A polytope is a bounded intersection of half-spaces [122]. The modes that are not realizable by a statistical model yield a (full dimensional) polyhedral approximation of their complement. This can be used to bound the approximation errors of the models from *below*. Indeed, the faces of $\mathcal{G}_n$ are portions of convex exponential families [79]. Maximizing the KL-divergence to these models is significantly easier than to general exponential families, not to mention mixtures. Strong modes are easier to study than not-strong modes, since they are described by fewer inequalities.

Example 1.B.3. Each p in the set $\mathcal{G}_{3,4}$ of probability distributions on $\{0,1\}^3$ which have four modes satisfies one of the following two sets of inequalities:

$$p(x) > p(y) \quad \forall y \in \{0,1\}^3 \text{ with } d_H(x,y) = 1 \quad \forall x \in Z_{+,3} \ \vee \ \forall x \in Z_{-,3}\ . \tag{1.12}$$

Either set of inequalities represents an intersection of open half-spaces, i.e., a (possibly unbounded) open h-polytope. The closure of the intersection of this h-polytope with the probability simplex is a bounded convex polytope $\overline{\mathcal{G}_3^{\pm}}$, and $\overline{\mathcal{G}_3} = \overline{\mathcal{G}_3^+} \cup \overline{\mathcal{G}_3^-}$. The v-presentation of a bounded polytope is a finite list of points in the polytope which contains the vertex set of the polytope. Table 1.1 shows the list of vertices of the closure of $\mathcal{G}_3^+$, the set of all probability distributions on $\{0,1\}^3$ which have four modes $Z_{+,3}$. The set of vertices incident to each of the facets of $\overline{\mathcal{G}_3^+}$ is listed in Table 1.2. The Lebesgue volume of the set $\mathcal{G}_3$ can be easily computed (using

$$\begin{array}{c}
\left(\begin{array}{cccccccc}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
\hline
0 & 1/4 & 1/4 & 1/4 & 0 & 0 & 0 & 1/4 \\
1/4 & 0 & 1/4 & 1/4 & 0 & 0 & 1/4 & 0 \\
1/4 & 1/4 & 0 & 1/4 & 0 & 1/4 & 0 & 0 \\
1/4 & 1/4 & 1/4 & 0 & 1/4 & 0 & 0 & 0 \\
\hline
1/6 & 1/6 & 1/6 & 1/6 & 0 & 1/6 & 0 & 1/6 \\
1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 0 & 0 & 1/6 \\
1/6 & 1/6 & 1/6 & 1/6 & 0 & 0 & 1/6 & 1/6 \\
1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 0 & 0 \\
1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 0 & 1/6 & 0 \\
1/6 & 1/6 & 1/6 & 1/6 & 0 & 1/6 & 1/6 & 0 \\
\hline
1/7 & 1/7 & 1/7 & 1/7 & 0 & 1/7 & 1/7 & 1/7 \\
1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 & 1/7 & 1/7 \\
1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 & 1/7 \\
1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\
\hline
1/8 & 1/8 & 1/8 & 1/8 & 1/8 & 1/8 & 1/8 & 1/8
\end{array}\right) \\
\begin{array}{cccccccc}
(000) & (011) & (101) & (110) & (001) & (010) & (100) & (111)
\end{array}
\end{array} \tag{1.14}$$

Table 1.1: *This table shows the v-presentation of the polytope $\overline{\mathcal{G}_3^+}$. Each row of the matrix is a probability distribution on $\{0,1\}^3$ which is a vertex of the polytope. The vertices in the first group are the point measures on $Z_{+,3}$. They have degree 18 (i.e., they are incident to 18 edges) and are connected by edges to any other vertex of the polytope. The vertices in the second and fourth groups have degree 11. There are no edges between pairs of vertices from the second group. The vertices in the third group have degree 8, and the uniform distribution has degree 12. The f-vector of the polytope is $(19, 110, 290, 387, 270, 96, 16)$. The volume is $\operatorname{vol}(\mathcal{G}_3^+)/\operatorname{vol}(\mathcal{P}_3) = (1/2257920)/(1/40320) = 0.0179...$. The vertices of $\overline{\mathcal{G}_3^\pm}$ are the extreme points of the set of uniform distributions on subsets of $\{0,1\}^3$ which can be covered by three disjoint faces of the unit cube.*

`Polymake`):

$$\operatorname{vol}(\mathcal{G}_3^+ \cup \mathcal{G}_3^-)/\operatorname{vol}(\mathcal{P}_3) = 2 \cdot (1/2257920)/(1/40320) = 2 \cdot 0.01785714286... \,. \tag{1.13}$$

Example 1.B.4. The set of probability distributions on $\{0,1\}^3$ with modes (110), (101) and (011), denoted $\mathcal{G}(\{(110),(101),(011)\})$, is described by

$$p(110) > p(111), p(100), p(010)$$

and analogous inequalities for $p(101)$ and $p(011)$. The closure of the set of points in $\mathcal{P}_3$ satisfying these inequalities is a polytope with the vertices shown in Table 1.3. The set of probability distributions on $\{0,1\}^3$ which have three modes is the union of all sets of the form $\mathcal{G}(\mathcal{Y})$, where $\mathcal{Y}$ is a binary code of minimum distance two and cardinality three.

2	3	4		5	8		10	12	13	16	17	18	19
2	3	4		5	7		9	12	14	15	17	18	19
2	3	4		5	6		11	13	14	15	16	18	19
1	3	4		6	8		10	12	13	16	17	18	19
1	3	4		6	7		9	12	14	15	17	18	19
1	3	4		5	6		9	10	11	15	16	17	19
1	2	4		7	8		10	12	13	16	17	18	19
1	2	4		6	7		11	13	14	15	16	18	19
1	2	4		5	7		9	10	11	15	16	17	19
1	2	3		7	8		9	12	14	15	17	18	19
1	2	3		6	8		11	13	14	15	16	18	19
1	2	3		5	8		9	10	11	15	16	17	19
1	2	3	4	6	7	8	12	13	14	18			
1	2	3	4	5	6	7	9	11	14	15			
1	2	3	4	5	6	8	10	11	13	16			
1	2	3	4	5	7	8	9	10	12	17			

(1.15)

Table 1.2: *This table shows the vertex-facet incidence of the polytope* $\overline{\mathcal{G}_3^+}$*. Each row gives a list of vertices incident to one facet of* $\overline{\mathcal{G}_3^+}$*. The number of each vertex is the number of the row in which it appears in Table 1.1. The set of vertices contained in any face of the polytope is given by an intersection of the* 16 *sets listed above. There are edges between vertices from the first and the second groups shown in Table 1.1 which represent probability distributions with supports of cardinality* 5*. The support of these distributions can't be packed by three faces of the cube. These distributions are not contained in* $\mathrm{Mixt}^3(\overline{\mathcal{E}_3^1})$*.*

$$\begin{array}{cccccccc}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\
0 & 1/3 & 0 & 1/3 & 0 & 1/3 & 0 & 0 \\
0 & 1/3 & 1/3 & 0 & 1/3 & 0 & 0 & 0 \\
\hline
0 & 1/4 & 1/4 & 1/4 & 0 & 0 & 0 & 1/4 \\
\hline
0 & 1/5 & 1/5 & 1/5 & 0 & 1/5 & 0 & 1/5 \\
0 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 0 & 1/5 \\
0 & 1/5 & 1/5 & 1/5 & 0 & 0 & 1/5 & 1/5 \\
0 & 1/5 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 0 \\
0 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 1/5 & 0 \\
0 & 1/5 & 1/5 & 1/5 & 0 & 1/5 & 1/5 & 0 \\
\hline
0 & 1/6 & 1/6 & 1/6 & 0 & 1/6 & 1/6 & 1/6 \\
0 & 1/6 & 1/6 & 1/6 & 1/6 & 0 & 1/6 & 1/6 \\
0 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 0 & 1/6 \\
0 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 0 \\
\hline
0 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 \\
(000) & (011) & (101) & (110) & (001) & (010) & (100) & (111)
\end{array} \tag{1.16}$$

Table 1.3: *Each row of the matrix shown above is a vertex of the polytope of probability distributions on* $\{0,1\}^3$ *with modes* (110)*,* (101) *and* (011)*.*

The set of probability distributions which have the maximal possible number of modes has an easy description as the disjoint union of two convex polytopes. The following lemma provides a description of the set of probability distributions on $\{0,1\}^n$ with maximal number of modes. For the sake of simplicity and generality, the formulation of this lemma includes the description of functions on $\{0,1\}^n$ with maximal number of modes, but which are not necessarily normalized or non-negative.

Consider the set $\mathbb{R}^{2^n}$ of real-valued functions on $\{0,1\}^n$. We call $x \in \{0,1\}^n$ a mode of the function $f \in \mathbb{R}^{2^n}$ if $f(x) > f(y)$ for all y with $d_H(x,y) = 1$. Any two modes of f have Hamming distance at least two. Hence 2^{n-1} is the maximal number of modes of any function on $\{0,1\}^n$. Denote $\tilde{\mathcal{G}}_{n,2^{n-1}} \equiv \tilde{\mathcal{G}}_n$ the set of functions on $\{0,1\}^n$ which have 2^{n-1} modes. If a function has 2^{n-1} modes, then the set of modes is $Z_{+,n}$ or $Z_{-,n}$. The set $\tilde{\mathcal{G}}_n$ is the union of the set of functions with modes $Z_{+,n}$ and the set of functions with modes $Z_{-,n}$. The set of functions with modes $Z_{\pm,n}$ is

$$\tilde{\mathcal{G}}_n^{\pm} = \{p \in \mathbb{R}^{2^n} : p(x) > p(y)\ \forall x \in Z_{\pm,n}\ \forall y \text{ with } d_H(x,y) = 1\}\ . \tag{1.17}$$

Every inequality $p(x) > p(y)$ defines an open half-space of $\mathbb{R}^{2^n}$, and the set $\tilde{\mathcal{G}}_n^{\pm}$ is an intersection of $n2^{n-1}$ open half-spaces.

The set of probability distributions on $\{0,1\}^n$ which have 2^{n-1} modes is just $\mathcal{G}_{n,2^{n-1}} \equiv \mathcal{G}_n = \tilde{\mathcal{G}}_n \cap \overline{\mathcal{P}_n}$. Hence $\mathcal{G}_n^+$ is the intersection of the $n2^{n-1}$ open half-spaces defining $\tilde{\mathcal{G}}_n^+$, the (2^n-1)-dimensional affine space $\mathrm{aff}(\mathcal{P}_n) = \{f \in \mathbb{R}^{2^n} : \sum_x f(x) = 1\}$, and the 2^{n-1} closed

half-spaces defined through $p(y) \geq 0\ \forall y \in Z_{\mp,n}$, (the inequalities $p(x) \geq 0\ \forall x \in Z_{\pm,n}$ are also satisfied when $p \in \tilde{\mathcal{G}}_n$).

Lemma 1.B.5.

(i) The set $\mathcal{G}'_n := \tilde{\mathcal{G}}_n \cap \operatorname{aff}(\mathcal{P}_n)$ is an affine double-cone. More precisely, it is the disjoint union of two open, affine, polyhedral, convex cones $\mathcal{G}'^{+}_n$ and $\mathcal{G}'^{-}_n$ with common apex $u = (1/2^n, \dots, 1/2^n)$, and which are the image of each other by the reflection through u.

(ii) The set $\tilde{\mathcal{G}}_n \subset \mathbb{R}^{2^n}$ has a non-empty interior and full dimension 2^n for any $n \in \mathbb{N}$. The sets $\mathcal{G}'_n \subset \operatorname{aff}(\mathcal{P}_n)$ and $\mathcal{G}_n \subset \overline{\mathcal{P}_n}$ have a non-empty interior and full dimension $2^n - 1$ for any $n \in \mathbb{N}$.

(iii) The sets $\mathcal{G}'^{+}_n$ and $\mathcal{G}'^{-}_n$ are separated by any affine hyperplane which intersects u and has a normal vector in the following cone:

$$\mathcal{N} := \Big\{ \sum_{x \in Z_{+,n}} \sum_{y : d_H(x,y)=1} \lambda_{(x,y)}(\delta_x - \delta_y) \colon \lambda_{x,y} \geq 0, \sum \lambda_{x,y} > 0 \Big\} ,$$

for example, by the affine hyperplane $\{u_{Z_{+,n}} - u_{Z_{-,n}}\}^{\perp} + u$. The union of these hyperplanes equals the complement of $\mathcal{G}'_n$. The set $\overline{\mathcal{G}'^{\pm}_n}$ doesn't contain any line, i.e., $\overline{\mathcal{G}'^{\pm}_n}$ is a salient *cone. In particular, $\overline{\mathcal{G}'^{+}_n} \cap \overline{\mathcal{G}'^{-}_n} = \{u\}$.*

(iv) Any line $\mathcal{L} = \{p + \lambda(p - q) \colon \lambda \in \mathbb{R}\} \subset \operatorname{aff}(\mathcal{P}_n)$ which intersects both $\mathcal{G}'^{+}_n$ and $\mathcal{G}'^{-}_n$, intersects their boundaries exactly once: $|\mathcal{L} \cap (\overline{\mathcal{G}'^{\pm}_n} \setminus \mathcal{G}'^{\pm}_n)| = 1$.

(v) If the convex hull of a set of points $\{p^i \in \operatorname{aff}(\mathcal{P}_n)\}_{i=1}^{k}$ intersects $\mathcal{G}'^{\pm}_n$, then each of the $n2^{n-1}$ inequalities defining $\mathcal{G}'^{\pm}_n$ is satisfied by at least one of the points p^i.

Proof.

(i) The uniform distribution is contained in any hyperplane $\{p \colon p(x) = p(y)\}$ for any $x \neq y$, and hence it is contained in the closure of $\mathcal{G}'^{\pm}_n$. Any plane $\{f \in \operatorname{aff}(\mathcal{P}_n) \colon f(x) = f(y)\}$ for some $x, y \in \{0,1\}^n$ with $d_H(x,y) = 1$ separates $\mathcal{G}'^{+}_n$ and $\mathcal{G}'^{-}_n$, because any point $f \in \mathcal{G}'^{\pm}_n$ satisfies $f(x) \gtrless f(y)$. In particular, the two sets are disjoint. The sets $\mathcal{G}'^{\pm}_n$ are open affine convex cones, because they are an intersection of affine open half-spaces which contain the point u in their bounding planes.

Let $f = u + v$ be some vector in $\operatorname{aff}(\mathcal{P}_n)$. If $f \in \mathcal{G}'^{+}_n$, then $v(x) > v(y)$ for all $x \in Z_{+,n}$ and all y with $d_H(x,y) = 1$, and $\sum_x v(x) = 0$. The vector $-v$ satisfies $-v(x) < -v(y)$ for all $y \in Z_{-,n}$ and x with $d_H(x,y) = 1$ and $\sum_x -v(x) = 0$. Therefore, $u - v \in \mathcal{G}'^{-}_n$, and $\mathcal{G}'^{-}_n$ contains the reflection of any point in $\mathcal{G}'^{+}_n$ through u.

(ii) For any $\mathcal{Y} \subseteq \{0,1\}^n$ let $u_{\mathcal{Y}}$ be the probability vector defined by $u_{\mathcal{Y}}(x) := 1/|\mathcal{Y}|$ if $x \in \mathcal{Y}$ and $u_{\mathcal{Y}}(x) = 0$ else. The open set $\tilde{\mathcal{G}}_n$ contains a neighborhood of $u_{Z_{\pm,n}}$ and therefore has a non-empty interior and full dimension. The claim for the sets $\mathcal{G}'_n$ and $\mathcal{G}_n$ follows from similar arguments.

(iii) For any pair $f, g \in \mathbb{R}^{2^n}$ denote $\langle f, g \rangle := \sum_{x \in \{0,1\}^n} f(x)g(x)$ the standard scalar product. Let d be any vector in $\mathcal{N}$. From the definition of $\mathcal{N}$ we have $\langle d, (1, \dots, 1) \rangle = \langle d, u \rangle = 0$. Any point $f \in \mathcal{G}'^{+}_n$ satisfies $f = u + v$ with $\|v\|_1 \neq 0$ and $\operatorname{sgn} v(x) \in \{0, +\}$ if $x \in Z_{+,n}$

and $\operatorname{sgn} v(x) \in \{0, -\}$ if $x \in Z_{-,n}$. Hence, $\langle v, d\rangle > 0$. On the other hand, any point w on the hyperplane with normal vector d, passing through u, satisfies $\langle d, (w-u)\rangle = 0$ and $\langle d, w\rangle = 0$. Hence $w \neq f$. Furthermore, the set of hyperplanes through u with normal vectors in $\mathcal{N}$ contains all bounding hyperplanes of the half-spaces defining $\mathcal{G}'^{+}_n$ and $\mathcal{G}'^{-}_n$.

The matrix with rows $\{(\delta_x - \delta_y)\}_{x \in Z_{+,n}, d_H(x,y)=1}$ has a one-dimensional kernel spanned by u. To see this note that any f in the kernel of the row $(\delta_x - \delta_y)$ satisfies $f(x) = f(y)$, and since the graph of the n-cube is Hamiltonian, this implies $f(x^1) = f(y^1) = \cdots = f(x^{2^n-1}) = f(y^{2^n-1})$ for $\{x^i\} = Z_{+,n}$ and $\{y^i\} = Z_{-,n}$. Hence the polytope $\operatorname{conv}\{(\delta_x - \delta_y)\}_{x \in Z_{+,n}, d_H(x,y)=1}$ has dimension $2^n - 2$. This implies that $\mathcal{G}'^{+}_n$ doesn't contain any line (the double-cone $\overline{\mathcal{G}'_n}$ certainly does), and $\overline{\mathcal{G}'^{+}_n} \cap \overline{\mathcal{G}'^{-}_n} = u$.

(iv) The generatrix of $\mathcal{G}'^{+}_n$ is the set of directions from the apex u to any other point in the boundary of $\mathcal{G}'^{+}_n$. Assume that the line $\mathcal{L}$ intersects the boundary of $\mathcal{G}'^{\pm}_n$ at two different points f and f'. Then the tangent vectors of $\mathcal{L}$ are not within the generatrix of $\mathcal{G}'^{\pm}_n$ and $f, f' \neq u$. Since $\mathcal{G}'^{\mp}_n$ is convex, a translate $\mathcal{G}'^{\mp}_n + p$ contains $\mathcal{G}'^{\mp}_n$ iff $p \in u - \mathcal{G}'^{\pm}_n$. Hence $\mathcal{G}'^{\mp}_n + (u - f) \supseteq \mathcal{G}'^{\mp}_n$. Since the tangent vectors of $\mathcal{L}$ are not within the generatrix of $\mathcal{G}'^{\pm}_n$, they are not within the generatrix of $\mathcal{G}'^{\mp}_n$ and $\mathcal{L}$ intersects the apex f of $\mathcal{G}'^{\mp}_n + (u - f)$, but $\mathcal{L}$ does not intersect any other point in $\overline{\mathcal{G}'^{\mp}_n} + (u - f)$, and in particular it does not intersect $\mathcal{G}'^{\mp}_n$. Hence if $\mathcal{L}$ intersects $\mathcal{G}'^{+}_n$ and $\mathcal{G}'^{-}_n$, then $|\mathcal{L} \cap \overline{\mathcal{G}'^{\pm}} \setminus \mathcal{G}'^{\pm}| \leq 1$. Since $\mathcal{G}'^{\pm}_n$ contains no line, any line which intersects $\mathcal{G}'^{\pm}_n$, intersects its boundary. This completes the proof.

(v) If not, the polytope lies in a half-space contained in the complement of $\mathcal{G}'^{\pm}_n$. □

Modes of Mixtures of Binary Independence Models

Definition 1.B.6. Let $1 \leq l \leq 2^{n-1}$. We define $g(l, n)$ to be the smallest $k \in \mathbb{N}$ for which there are k points in $\mathcal{E}^1_n$ whose convex hull contains a distribution with l modes. Similarly, we define $h(l, n)$ to be the smallest $k \in \mathbb{N}$ for which there are k points in $\mathcal{E}^1_n$ whose convex hull contains a distribution with l strong modes.

Question 1.B.7. What is $h(l, n)$, $g(l, n)$?

Corresponding questions have been posed in the case of continuous variables and mixtures of multivariate normal distributions [97]. At the present time the maximal number of modes of a mixture of k normal distributions on $\mathbb{R}^n$ is unknown.

The following lemma extends Corollary 1.3.3. It gives inequalities which cut out a portion of the complement of the mixture model $\operatorname{Mixt}^k(\mathcal{E}^1_n)$.

Lemma 1.B.8. *Let $n \in \mathbb{N}$ and $p \in \overline{\mathcal{P}_n}$. If $x \in \{0,1\}^n$ is a strong mode of p, then any representation of p as mixture of product distributions includes a mixture component which is strictly maximized at x. Furthermore, the smallest k for which $\operatorname{Mixt}^k(\mathcal{E}^1_n)$ contains a distribution with l strong modes is*

$$h(l, n) = l \quad \forall l \in \{0, \ldots, 2^{n-1}\} ,$$

and

$$\left(\overline{\mathcal{P}_n} \setminus \operatorname{Mixt}^k(\overline{\mathcal{E}^1_n})\right) \supseteq \mathcal{H}_{n,k+1} \quad \forall k .$$

Proof. Consider any $p \in \overline{\mathcal{E}^1_n}$. The set of maximizers of p is a face of C_n and hence p has at most one mode. Consider a pair of vectors $x, y \in \{0,1\}^n \cong \mathbb{F}_2^n$. We can write $y = x +$

$\mathbb{1}_{\mathrm{supp}(x)\triangle\mathrm{supp}(y)} (\mathrm{mod}\, 2)$, where $\triangle$ denotes the symmetric difference. If $p(x) \geq p(y)$, then $p(x) \geq p(z) \geq p(y)$ for any $z = x + \mathbb{1}_B$ with $B \subseteq \mathrm{supp}(x) \triangle \mathrm{supp}(y)$. Consider now a probability distribution q which has a strong mode x. We show that any mixture decomposition of q into product distributions has a mixture component with mode x. Let q^i, $i \in [n]$, denote the mixture components. Assume that non of them has mode x. Then, for every i, there exists a y such that $q^i(y) \geq q^i(x)$, and $q^i(\hat{x}) \geq q^i(x)$ for $\hat{x} = x + \mathbb{1}_{\{j\}}$ for any $j \in \mathrm{supp}(x) \triangle \mathrm{supp}(y)$. Obviously, $d_H(\hat{x}, x) = 1$, but this implies that x can't be a strong mode of $\sum_i \alpha_i q^i$.

Corollary 1.3.3 implies that $\mathrm{Mixt}^k(\overline{\mathcal{E}_n^1})$ contains any distribution with support of cardinality k; in particular the uniform distribution on a binary code of minimum distance two and cardinality k whenever $k \leq 2^{n-1}$. This is a distribution from $\mathcal{H}_{n,k}$. □

Corollary 1.B.9. *The model* $\mathrm{Mixt}^k(\mathcal{E}_n^1)$ *intersects* $\mathcal{H}_{n,m}$ *if and only if* $\mathrm{Mixt}^k(\overline{\mathcal{E}_n^1})$ *contains a distribution supported by a binary code of minimum distance two and cardinality at least* m.

The following is a crude bound for the Lebesgue volume of the set of probability distributions with more than m strong modes and by Lemma 1.B.8, for the complement of $\mathrm{Mixt}^m(\mathcal{E}_n^1)$:

Proposition 1.B.10. *If* $m < 2^{n-1}$, *then* $\mathrm{vol}(\mathcal{H}_{n,m+1}) \geq K(m+1) 2^{-(m+1)n} \mathrm{vol}(\mathcal{P}_n)$, *where*

$$K(m+1) = \begin{cases} 2^{m+1}, & \text{if } m+1 \leq 2^k \leq \frac{2^n}{n} \text{ for some } k \\ 2, & \text{otherwise} \end{cases}.$$

Proof. (i) The probability simplex $\overline{\mathcal{P}}$ is a regular $(|\mathcal{X}|-1)$-simplex in $\mathbb{R}^{|\mathcal{X}|}$; all edges have the same length $\sqrt{2}$. Let $\mathcal{H}(\mathcal{Y})$ denote the set of probability distributions which have strong modes at $\mathcal{Y} \subset \mathcal{X}$. Then $\mathcal{H}(\mathcal{Y}) = \cap_{y \in \mathcal{Y}} \mathcal{H}(y)$. For any $y \in \mathcal{X}$, denote by $B_1(y)$ the Hamming ball with center y and radius 1. The set $\overline{\mathcal{P}}_y(B_1(y)) := \{p \in \overline{\mathcal{P}}(B_1(y)) \colon p(y) \geq p(\mathcal{X} \setminus \{y\})\}$ is a regular n-simplex of side length $\frac{\sqrt{2}}{2}$ (its vertices are $\{\frac{1}{2}(\delta_y + \delta_{\hat{y}})\}_{d_H(\hat{y},y)\leq 1}$). The volume of a regular N-simplex with side length l is $\mathrm{vol}(\Delta_l^N) = \frac{\sqrt{N+1}}{N! \sqrt{2}^N} l^N$. The set $\mathcal{H}(y)$ is the convex hull of $\overline{\mathcal{P}}_y(B_1(y))$ and $\overline{\mathcal{P}}(\mathcal{X} \setminus B_1(y))$, and hence $\mathrm{vol}\, \mathcal{H}(y) = 2^{-n} \mathrm{vol}\, \mathcal{P}$. If $\mathcal{Y}$ has minimum distance 3 or more, $\mathrm{vol}\, \mathcal{H}(\mathcal{Y}) = 2^{-|\mathcal{Y}| \cdot n} \mathrm{vol}\, \mathcal{P}$. If the minimum distance of $\mathcal{Y}$ is two, then the volume of $\mathcal{H}(\mathcal{Y})$ is larger.

(ii) The number $K(m+1)$ is a lower bound on the number of disjoint sets $\mathcal{H}(\mathcal{Y})$ with $|\mathcal{Y}| = m+1$. The Gilbert-Varshamov bound (see Proof of Theorem 1.3.1) tells us that if $m+1 \leq 2^k$, where k is the largest integer for which $2^k \leq \frac{2^n}{n}$, there exists a binary code $\mathcal{Y} \subset \mathcal{X}$, $|\mathcal{Y}| = m+1$ with minimal distance 3. Let $\mathcal{Y}' = \mathcal{Y} \setminus \{y\} \cup \{y + \mathbb{1}_1\}$ (flip one coordinate of one element of $\mathcal{Y}$). We have that $\mathcal{H}(\mathcal{Y}) \cap \mathcal{H}(\mathcal{Y}') = \emptyset$. Since $\mathcal{Y}$ has $(m+1)$ elements, there are 2^{m+1} disjoint sets $\mathcal{H}$. For any $m+1 \leq 2^{n-1}$, if $\mathcal{Y}$ is a binary code of minimum distance 2, then also $\mathcal{Y}' + \mathbb{1}_1$, and $\mathcal{H}(\mathcal{Y}) \cap \mathcal{H}(\mathcal{Y}') = \emptyset$. □

Remark 1.B.11.

(i) There are mixtures of two product distributions which have more than two modes. Hence $g(l,n) \neq h(l,n)$, in general.

(ii) The set of distributions with l strong modes is contained in the set of distributions with l modes. By Lemma 1.B.8 we get

$$g(l,n) \leq h(l,n) = l \quad \forall l \in \{0, \ldots, 2^{n-1}\}\,. \tag{1.18}$$

The following lemma relates the number of modes of a mixture model to the number of modes of *truncations* of the model:

Lemma 1.B.12. *Let $n \geq 2$ and $k \in \mathbb{N}$. Let p be a non-negative sum of k products $p = \sum_{i\in[k]} \lambda_i \prod_{j\in[n]} p^{i,j}$, with $\lambda_i \geq 0$ and $(p^{i,1}, \ldots, p^{i,n}) \in (\mathcal{P}_1)^n$, $i = 1, \ldots, k$. If p has 2^{n-1} modes, then for any $\{j_1, \ldots, j_m\} \subsetneq \{1, \ldots, n\}$ the convex hull of the products $\prod_{l\in[m]} p^{i,j_l} \in \mathcal{E}^1_m$, $i = 1, \ldots, k$ on $\{0,1\}^{j_1,\ldots,j_m} \cong \{0,1\}^m$ intersects both $\mathcal{G}^+_m$ and $\mathcal{G}^-_m$.*

Proof. We show an illustrative special case of the claim. The proof of the general case is a straightforward generalization of the proof of this special case:

Let $n \geq 2$ and $k \in \mathbb{N}$. If $q \in \mathrm{Mixt}^k(\mathcal{E}^1_n)$ is contained in $\mathcal{H}^+_n$, then there exist k elements of $\mathcal{E}^1_{n-1}$ whose convex hull intersects both $\mathcal{G}^+_{n-1}$ and $\mathcal{G}^-_{n-1}$.

Any mixture of k product distributions with n variables, $q \in \mathrm{Mixt}^k(\mathcal{E}^1_n)$, has the following form:

$$q(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{k} \lambda_i p^{i,1}(x_1) p^{i,2}(x_2) \cdots p^{i,n}(x_n) \ , \quad \forall (x_1, x_2, \ldots, x_n) \in \{0,1\}^n \ , \tag{1.19}$$

where $\sum_{i=1}^k \lambda_i = 1$, $\lambda_i \geq 0$ and $p^{i,j} \in \mathcal{P}_1$. For the fixed value $x_1 = 0$ the above expression can be understood as a mixture of k product distributions with $(n-1)$ variables, multiplied by a positive constant:

$$q(x_1 = 0, x_2, \ldots, x_n) = c_0 \sum_{i=1}^{k} \lambda_{0,i} p^{i,2}(x_2) \cdots p^{i,n}(x_n) = c_0 q_0(x_2, \ldots, x_n) \ , \\ \forall (x_2, \ldots, x_n) \in \{0,1\}^{n-1} \ , \tag{1.20}$$

where $\sum_{i=1}^k \lambda_{0,i} = 1, \lambda_{0,i} \geq 0$ with

$$\lambda_{0,i} = \frac{\lambda_i p^{i,1}(x_1 = 0)}{c_0} \quad \text{and} \quad c_0 = \sum_{i=1}^{k} \lambda_{0,i} p^{i,1}(x_1 = 0) \ . \tag{1.21}$$

A similar observation can be made for the fixed value $x_1 = 1$. In total we get the following:

$$q = \begin{cases} c_0 \sum_{i=1}^k \lambda_{0,i} p^{i,2} \cdots p^{i,n} = c_0 q_0 \ , & \text{if } x_1 = 0 \\ c_1 \sum_{i=1}^k \lambda_{1,i} p^{i,2} \cdots p^{i,n} = c_1 q_1 \ , & \text{if } x_1 = 1 \end{cases} . \tag{1.22}$$

If the probability distribution q is contained in $\mathcal{G}^+_n$, then

(i) $q_0 \in \mathcal{G}^+_{n-1}$, which is to say that q satisfies the inequalities describing $\mathcal{G}^+_n$ involving coordinates from the set $\{(0, x_2, \ldots, x_n) \colon (x_2, \ldots, x_n) \in \{0,1\}^{n-1}\}$.

(ii) $q_1 \in \mathcal{G}^-_{n-1}$, which is to say that q satisfies the inequalities describing $\mathcal{G}^+_n$ involving coordinates from the set $\{(1, x_2, \ldots, x_n) \colon (x_2, \ldots, x_n) \in \{0,1\}^{n-1}\}$.

The probability distributions q_0 and q_1 are mixtures of the same k product distributions $\{p^{i,2} \cdots p^{i,n} \in \mathcal{E}^1_{n-1}\}_{i=1,\ldots,k}$, although they may have different mixture weights $(\lambda_{0,1}, \ldots, \lambda_{0,k})$ and $(\lambda_{1,1}, \ldots, \lambda_{1,k})$. The convex hull of $\{\prod_{j=2}^n p^{i,j} \in \mathcal{E}^1_{n-1}\}_{i=1}^k$ contains q_0 and q_1, and intersects $\mathcal{G}^+_{n-1}$ and $\mathcal{G}^-_{n-1}$. □

Corollary 1.B.13. $g(2^n, n+1) > g(2^{n-1}, n)$ *for all* $n \geq 1$.

Proof. This follows from Lemma 1.B.12 and Lemma 1.B.5*(iv)*. □

The Three-Mixture of Three Independent Variables

Example 1.B.14. (Modes of $\mathrm{Mixt}^3(\mathcal{E}^1_{3,\mathrm{bin}})$). We have seen that any element of $\mathrm{Mixt}^3(\mathcal{E}^1_{3,\mathrm{bin}})$ has at most three strong modes. What about not-strong modes? We ask whether the following inequalities have a solution:

$$\sum_{i=1}^{3} \alpha_i (p_i(x) - p_i(y)) > 0 \quad \forall y \text{ s.t. } d_H(x,y) = 1,\ \forall x \in Z_{+,3},\ \text{for } p_i \in \overline{\mathcal{E}^1_{3,\mathrm{bin}}}\,. \tag{1.23}$$

Every mixture component is a product $p_i(x) = p_i^1(x_1) \cdot p_i^2(x_2) \cdot p_i^3(x_3)$. Equation (1.23) represents a problem with 12 non-linear inequalities involving polynomials of degree four in the 11 variables $\{p_i^j\}_{i\in\{1,2,3\},j\in\{1,2,3\}}$ and α_1, α_2, and 22 linear inequality constraints $0 \leq p_i^j(x_j = 1) \leq 1$, and $0 \leq \alpha_i$, $\alpha_1 + \alpha_2 \leq 1$. We formulated this as constrained optimization problem with objective

$$f(x) = \prod_{x\in Z_+, y: d_H(x,y)=1} \Big(\sum_i \alpha_i (p_i(x) - p_i(y))\Big)^2\,, \tag{1.24}$$

run the MATLAB function `fmincon` on a large number of initial values and found no solutions. There exist solutions where the inequalities from (1.23) are not strict and the objective vanishes, e.g., the uniform distribution, or the point measures within $\overline{\mathcal{P}}(Z_+)$. An alternative approach is to search for likelihood maximizers within the mixture model given a target with four modes: We run a custom EM algorithm for (data generated from) targets of the form $p_{\lambda,\beta} = (1-\lambda)u + \lambda(\beta u_{Z_+} + (1-\beta)u_{\mathcal{Y}})$ with $\mathcal{Y} \subseteq Z_+$. The resulting likelihood maximizers had a strictly positive Kullback-Leibler divergence to the target whenever $\lambda, \beta > 0$. The result can be seen in Figure 1.4. Both numeric evaluations suggest that any $p \in \mathrm{Mixt}^3(\overline{\mathcal{E}^1_{3,\mathrm{bin}}})$ has at most three modes. With the next Proposition 1.B.15 we provide a rigorous proof of this statement.

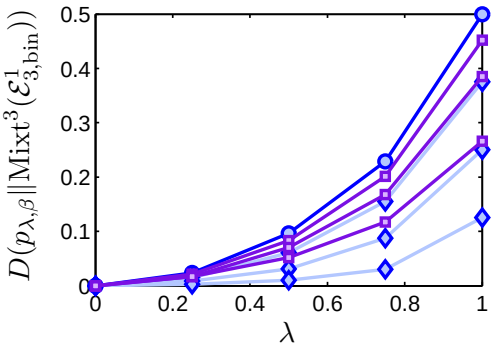

Figure 1.4: *This figure shows the Kullback-Leibler divergence from probability distributions on* $\{0,1\}^3$ *which have four modes, to the model* $\mathrm{Mixt}^3(\mathcal{E}^1_{3,\mathrm{bin}})$. *The line marked with circles shows* $p_\lambda = (1-\lambda)u + \lambda u_{Z_+}$, *the "symmetric" distributions with four modes on* Z_+. *The other lines show* $p_{\lambda,\beta} = (1-\lambda)u + \lambda(\beta u_{Z_+} + (1-\beta)u_{\mathcal{Y}})$ *for* $\beta \in \{0.25, 0.5, 0.75\}$*; the squares correspond to* $\mathcal{Y} = \{(000), (011), (110)\}$, *and the diamonds to both* $\mathcal{Y} = \{(000), (011)\}$ *and* $\mathcal{Y} = \{(000)\}$.

Proposition 1.B.15. *The mixture model consisting of convex combinations of any three product distributions on* $\{0,1\}^3$ *doesn't contain any probability distribution which has four modes:*

$$\mathrm{Mixt}^3(\overline{\mathcal{E}_3^1}) \cap \mathcal{G}_3 = \emptyset \ .$$

In particular u *is not an inner point of the model* $\mathrm{Mixt}^3(\overline{\mathcal{E}_3^1})$.

Proof. Assume that $\mathrm{Mixt}^3(\mathcal{E}_3^1) \cap \mathcal{G}_3 \neq \emptyset$. Without loss of generality assume $\mathrm{Mixt}^3(\mathcal{E}_3^1) \cap \mathcal{G}_3^+ \neq \emptyset$. By Lemma 1.B.12 there exist $(p^{i,1}, p^{i,2}, p^{i,3}) \in (\mathcal{P}_1)^3, i = 1,2,3$ such that the convex hull $\mathrm{conv}\{q^i\}_{i=1,2,3}$ of the product distributions $q^i := p^{i,2}p^{i,3} \in \mathcal{E}_2^1$ intersects $\mathcal{G}_2^+$ and $\mathcal{G}_2^-$. By Lemma 1.B.5*(iv)* any line which intersects both $\mathcal{G}_2'^+$ and $\mathcal{G}_2'^-$, intersects the boundary of each of them exactly once. See Figure 1.5. If a line segment in $\mathrm{conv}\{q^i\}_i$ extends to a line which intersects the boundaries of $\mathcal{G}_2'^+$ and $\mathcal{G}_2'^-$ exactly once, then the distributions q^1, q^2, q^3 can be enumerated such that $\mathrm{conv}\{q^1, q^2\}$ intersects $\mathcal{G}_2^+$, and $\mathrm{conv}\{q^2, q^3\}$ intersects $\mathcal{G}_2^-$. By Lemma 1.B.5*(v)* the mixture of $\mathrm{conv}\{q^2, q^3\}$ intersects $\mathcal{G}_2^-$ only if $q^2 \in \mathcal{G}(\{(10)\})$ and $q^3 \in \mathcal{G}(\{(01)\})$ or $q^3 \in \mathcal{G}(\{(10)\})$ and $q^2 \in \mathcal{G}(\{(01)\})$. Similarly, if $\mathrm{conv}\{q^1, q^2\}$ intersects $\mathcal{G}_2^+$, then $q^1 \in \mathcal{G}(\{(11)\})$ and $q^2 \in \mathcal{G}(\{(00)\})$ or $q^2 \in \mathcal{G}(\{(11)\})$ and $q^1 \in \mathcal{G}(\{(00)\})$. Contradiction! The uniform distribution is contained in the independence model and hence also $u \in \mathrm{Mixt}^3(\mathcal{E}_3^1)$. On the other hand, by Lemma 1.B.5, u is contained in the closure of $\mathcal{G}_3$. □

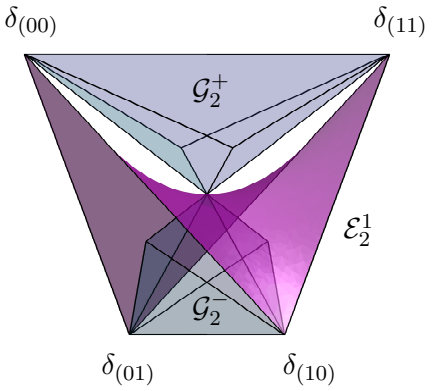

Figure 1.5: *This figure shows the* 3*-dimensional probability simplex on* $\{0,1\}^2$*, the independence model* $\mathcal{E}_2^1$*, and the sets of distributions with two modes* $\mathcal{G}_2^+$ *and* $\mathcal{G}_2^-$.

Remark 1.B.16. In the proof of Proposition 1.B.15 we used only 8 of the 12 inequalities that describe $\mathcal{G}_3^+$.

In the next chapters of this thesis we will study the class of binary models represented by Restricted Boltzmann Machines. Without going into details at this point, we discuss a particular example; the Restricted Boltzmann Machine model $\mathrm{RBM}_{4,2}$. This model is contained in $\mathrm{Mixt}^4(\overline{\mathcal{E}_4^1})$ and has codimension one in $\overline{\mathcal{P}_4}$. Its *algebraic implicitization* was studied in [32], i.e., its description as the set of zeros of a collection of polynomials. Using intensive computer analysis Cueto et al. [32] showed that such a description is very complicated and involves polynomials of degree 110 in as many as 5.5 trillion monomials. By Lemma 1.B.8 $\mathrm{Mixt}^4(\mathcal{E}_4^1) \cap \mathcal{H}_4 = \emptyset$ and $\mathrm{RBM}_{4,2} \cap \mathcal{H}_4 = \emptyset$. Proposition 1.B.15 and Corollary 1.B.13 allow us to explicitly describe a larger portion of the complement:

Corollary 1.B.17. $\mathrm{Mixt}^4(\mathcal{E}_4^1)$ *doesn't intersect* $\mathcal{G}_4$*. In particular, the Restricted Boltzmann Machine with four visible and two hidden units can't represent any probability distribution with eight modes*

$$\mathrm{RBM}_{4,2} \cap \mathcal{G}_4 = \emptyset \ .$$

In the following we show that the convex hull of certain submodels of $\mathcal{E}_n^1$ doesn't intersect $\mathcal{G}_n$.

For any $k \in \mathbb{N}$, $B_i \in \mathbb{R}^n$ and $\lambda_i \in \mathbb{R}_{\geq 0}$ for all $i = 1, \ldots, k$ we consider the following function:

$$F_{B_i,\lambda_i} \colon \mathbb{R}^n \to \mathbb{R} \ ; \quad x \mapsto \sum_{i=1}^{k} \lambda_i \exp(\sum_{j=1}^{n} B_{i,j} x_j) \ . \tag{1.25}$$

We will use the following elementary observation:

Proposition 1.B.18. *Let* $k, n \in \mathbb{N}$*. If* $\{B_i\}_{i\in[k]}$ *are any vectors in* $\mathbb{R}^n$*,* $\lambda_i \in \mathbb{R}_{\geq 0}$ *for all* $i \in [k]$*, and* F_{B_i,λ_i} *is defined as in eq.* (1.25)*, then* F_{B_i,λ_i} *is a convex function. If* $0 \in \mathrm{aff}\{B_i\}_{i\in[k]}$*, then* $\nabla_w F = 0$ *for all* $w \in \mathrm{aff}\{B_i\}^{\perp}$*. If* $\dim(\mathrm{aff}\{B_i\}_{i\in[k]}) < n$*, then each level set is unbounded.*

Proof. The Hessian $H(F)(x) = (\sum_i \lambda_i B_{i,j} B_{i,k} \exp(B_i x))_{jk}$ is everywhere positive semidefinite, since

$$y^\top (H(F)(x))y = \sum_i \lambda_i (\sum_j y_j B_{i,j})(\sum_k y_k B_{i,k}) \exp(B_i x))_{jk} = \sum_i \lambda_i \langle y, B_i \rangle^2 \exp(B_i x))_{jk} \ .$$

The Hessian is everywhere positive definite and the function F is strictly convex if and only if the vectors B_i span $\mathbb{R}^n$. □

From the convexity of F it follows that each sublevel set $L_c^-(F) := \{x \in \mathbb{R}^n \colon F_{B_i,\lambda_i}(x) \leq c\}$ is a convex subset of $\mathbb{R}^n$.

Proposition 1.B.19. *If* M *is a linear projection of* $\mathbb{R}^3$ *onto a* 2*-space, then* $\mathrm{conv}(M(Z_{+,3}))$ *intersects* $M(Z_{-,3})$*. Hence* $M(Z_{+,3})$ *and* $M(Z_{-,3})$ *can't be separated by the boundary of a convex set.*

Proof. Use that the convex hull of any two points in $M(Z_{-,3})$ intersects the convex hull of two points in $M(Z_{+,3})$ (this follows from Radon's theorem applied to any 2-face of $\{0,1\}^3$). If $\mathrm{conv}(M(Z_{+,3}))$ didn't contain one element of $M(Z_{-,3})$, then the above condition wouldn't hold. □

Corollary 1.B.20. *Let* $\mathcal{E}$ *be any two-dimensional exponential subfamily of* $\mathcal{E}_3^1$ *which contains* u*. Then* $\mathrm{conv}(\mathcal{E})$ *doesn't contain any probability distribution of the form*

$$p = \lambda u + (1-\lambda) u_{Z_{\pm,3}} \ , \quad \lambda \in [0,1) \ .$$

In particular, u *is not an inner point of* $\mathrm{conv}(\mathcal{E})$*.*

Proof. If $\mathcal{E}$ is a two-dimensional exponential subfamily of the independence model $\mathcal{E}_n^1$, then any element from $\mathcal{E}$ has the form $\exp(B^\top x)/Z$, where $x \in \{0,1\}^n$, and B belongs to a 2-dimensional subspace of $\mathbb{R}^n$, and Z is a normalization constant. Any element in a mixture of k

elements from $\mathcal{E}$ is proportional to

$$p(x) \propto F(x) = \sum_{i=1}^{k} \lambda_i \exp(B_i^\top x) \quad \forall x \in \{0,1\}^n \,. \tag{1.26}$$

Each level set of F bounds a convex set which is symmetric along the normal vectors of $\operatorname{aff}\{B_i\}$. Together with Proposition 1.B.19 this implies that $p \neq \lambda u + (1-\lambda) u_{Z_{\pm,3}}, \forall \lambda \in (0,1]$. □

Example 1.B.21. The models from Corollary 1.B.20 include the one-dimensional model of n independent and identically distributed variables $\{p_\theta(x) \propto \exp(\theta \mathbb{1}^\top x) \colon \theta \in \mathbb{R}\}$. This model is a curve of order n contained in the *exchangeable simplex* with vertices $\binom{n}{k}^{-1} \sum_{x, \|x\|_1 = k} \delta_x$ for $k = 0, \ldots, n$. For $n = 3$ the convex hull of this model doesn't contain any distribution with four modes.

1.C Hadamard Matrices and Related Exponential Families

In this appendix we study submatrices of Hadamard matrices. We are interested in the matrix with entries

$$\sigma(\lambda, x) := \prod_{r \in \lambda} x_r \,, \quad \text{for } \lambda \in 2^{[N]} \text{ and } x \in \mathcal{X} = \{-1,+1\}^N \,,$$

where x_r is the r-th coordinate of x. We use the following notation: Let σ be a matrix with rows indexed by $a \in A$, and columns indexed by $b \in B$. For any subsets $C \subseteq A$ and $D \subseteq B$ we write $\sigma(C, D) := (\sigma(a,b))_{a \in C, b \in D}$. For a given λ, $\sigma(\lambda, \cdot)$ is a function on $x \in \mathcal{X}$ called *character* which only depends on the values of x in the coordinates $i \in \lambda$. These functions build a basis of the space of functions on $\mathcal{X}$, and are used to define the sufficient statistics of binary hierarchical models (see Section 1.1). The matrix σ is a *Hadamard* matrix; it is a square $2^N \times 2^N$ matrix with entries ± 1 which satisfies $\sigma^\top \sigma = 2^N I$, where I is the identity matrix. All rows of σ are orthogonal to each other and have the same two-norm. The indices $\lambda \in 2^{[N]}$ and $x \in \mathcal{X}$ can be arranged in such a way that σ is symmetric. In that case σ is called a *Sylvester* matrix, and can be represented by the following Kronecker product $\sigma(2^{[N]}, \{\pm 1\}^N) = \otimes_{i=1}^N \binom{+1+1}{+1-1}$. See [60] for details.

We use the following nomenclature:

- For any $\Delta \subseteq 2^{[N]}$ let $\mathcal{Y}_\Delta := \{x \in \mathcal{X} : \operatorname{supp}(\mathbb{1} - x) \in \Delta\}$. In particular, the set $\mathcal{Y}_{\Delta_k}^{(y)}$ consists of all elements $z \in \mathcal{X}$ with Hamming distance to y smaller or equal to k.
- For any $\mathcal{Y} \subseteq \mathcal{X}$ let $\mathcal{Y}^{[-]} := \{x : -x \in \mathcal{Y}\}$, and more generally $\mathcal{Y}^{[\xi]} := \{\xi_i x^{(i)} : \{x^{(i)}\}_i = \mathcal{Y}\}$ for any $\xi = (\xi_1, \ldots, \xi_{|\mathcal{Y}|}) \in \{\pm 1\}^{|\mathcal{Y}|}$.
- For any $\mathcal{Y} \subseteq \mathcal{X}$ and $y \in \mathcal{X}$ we define $\mathcal{Y}^{(y)} := \{z = x * y : x \in \mathcal{Y}\}$, where $*$ denotes the element-wise product $x * y := \operatorname{diag}(y) x$, where $\operatorname{diag}(y)$ is the diagonal matrix with entires y_i in its diagonal.
- The *radius* of $\mathcal{Y} \subseteq \mathcal{X}$ is $\min_{x \in \mathcal{Y}} \max_{x' \in \mathcal{Y}} |\operatorname{supp}(x - x')|$. Given any element $x \in \mathcal{X}$ the set of elements in $\mathcal{X}$ which have Hamming distance to x at most k is a set of radius k.

We present first the results on submatrices of Hadamard matrices, then their proofs, and at the end the applications to related exponential families.

Lemma 1.C.1.

(i) Let $\Delta = 2^\lambda \subseteq 2^{[N]}$. The sub-matrix $\sigma(\Delta, \mathcal{Y}_\Delta)$ has orthogonal rows.

(ii) Consider any $\Delta \subseteq 2^{[N]}$ and any $\mathcal{Y} \subseteq \mathcal{X}$. The matrix $\sigma(\Delta, \mathcal{Y})$ has full rank $\min\{|\mathcal{Y}|, |\Delta|\}$ iff $\sigma(\Delta^c, \mathcal{Y}^c)$ has full rank $\min\{|\mathcal{Y}^c|, |\Delta^c|\}$.

(iii) For any simplicial complex $\Delta \subseteq 2^{[N]}$ the matrix $\sigma(\Delta, \mathcal{Y}_\Delta)$ has full rank $|\Delta|$, and the matrix $\sigma(\Delta^c, \mathcal{Y}^c_\Delta)$ has full rank $|\Delta^c|$.

(iv) Consider any $\Delta \subseteq 2^{[n]}$ and $\mathcal{Y} \subseteq \mathcal{X}$. If $\operatorname{rk}\sigma(\Delta, \mathcal{Y}) = |\Delta|$, then

a) $\operatorname{rk}\sigma(\Delta, \mathcal{Y}^{[-]}) = |\Delta|$.

b) If Δ consists of sets of equal cardinality, then $\operatorname{rk}\sigma(\Delta, \mathcal{Y}^{[\xi]}) = |\Delta| \ \ \forall \xi \in \{\pm 1\}^{|\mathcal{Y}|}$.

(v) Consider any $\Delta \subseteq 2^{[n]}$ and $\mathcal{Y} \subseteq \mathcal{X}$. For any $y \in \mathcal{X}$ we have $\operatorname{rk}\sigma(\Delta, \mathcal{Y}) = R \Leftrightarrow \operatorname{rk}\sigma(\Delta, \mathcal{Y}^{(y)}) = R$. Furthermore, $\sigma(\Delta, y * x) = \operatorname{diag}(\sigma(\Delta, y))\sigma(\Delta, x)$, where $y * x := \operatorname{diag}(y)x$.

In general, $y * \mathcal{Y} =: \mathcal{Y}^{(y)}$ can be equal to $\mathcal{Y}$ for $y \neq \mathbb{1}$. However, in the following cases the sets $\mathcal{Y}$ and $\mathcal{Y}^{(y)}$ are different:

Proposition 1.C.2.

(i) If $k < N$, then $\mathcal{Y}^{(y)}_{\Delta_k} \neq \mathcal{Y}^{(y')}_{\Delta_k}$, for any $y, y' \in \mathcal{X}, y \neq y'$.

(ii) If $|\mathcal{Y}|$ is odd, then all $\mathcal{Y}^{(y)}$, $y \in \mathcal{X}$ are different.

(iii) If $|\mathcal{Y}|$ is even, then there are at least $\frac{2^N}{|\mathcal{Y}|}$ different $\mathcal{Y}^{(y)}$, $y \in \mathcal{X}$.

The following Lemma 1.C.3 is similar to a lemma due to Alon [3] (see [63, Chapter 15]):

Lemma 1.C.3. *Let $v^i = (v^i_1, \ldots, v^i_n), i = 1, \ldots, k$ be k mutually orthogonal vectors in $\{\pm 1\}^n$, all of them orthogonal to $(1, \ldots, 1)$, and let $\alpha = (\alpha_1, \ldots, \alpha_k)$ be a vector in $\mathbb{R}^k \setminus \{(0, \ldots, 0)\}$. Let $v = \sum_{i=1}^k \alpha_i v^i$ and $v^+_j = \max\{0, v_j\}$. Then $v^+ := (v^+_1, \ldots, v^+_n)$ has more than $n/4k$ nonzero entries.*

We will use the following, which is a corollary of the results by Alon mentioned above. Its proof can be found in [63, Chapter 15]:

Corollary 1.C.4. *The sub-matrix given by $\sigma(A, B)$, where $A \subseteq 2^{[N]}$ and $B \subseteq \mathcal{X}$ and $|A| =: r$, $|B| =: t$ has rank r whenever $t > \left(1 - \frac{1}{r}\right)|\mathcal{X}|$.*

Proof of Lemma 1.C.1. Item (i) Consider any $x, y \in \mathcal{X}$ and let $\lambda' = \operatorname{supp}(\mathbb{1} - x) \triangle \operatorname{supp}(\mathbb{1} - y)$, such that $\lambda' \cap \Lambda = \emptyset$ iff $x_\Lambda = y_\Lambda$ for any $\Lambda \subset [N]$. We have the following:

$$\langle \sigma(2^\Lambda, x), \sigma(2^\Lambda, y)\rangle = \sum_{\lambda \in 2^\Lambda} \sigma(\lambda, x)\sigma(\lambda, y) = \sum_{\lambda \in 2^\Lambda} (-1)^{|\lambda' \cap \lambda|}$$
$$= \sum_{\lambda''' \subseteq \Lambda \setminus \lambda'} \ \sum_{\lambda'' \subseteq \lambda' \cap \Lambda} (-1)^{|\lambda''|} = 2^{|\Lambda|}\delta_{x_\Lambda, y_\Lambda} \ .$$

In the last equality we used that any $\Lambda \neq \emptyset$ has an equal number of subsets of even and odd cardinalities. Here the empty-set has even cardinality.

Item (ii) Consider first the case $|\mathcal{Y}| = |\Delta|$. It suffices to show one direction, since one may define $\Delta' = \Delta^c, \mathcal{Y}' = \mathcal{Y}^c$. Since $\sigma(\Delta, \mathcal{Y})$ has full rank $|\Delta|$, to every $z \in \mathcal{Y}^c$ there exists a vector $\tilde{v}_z = \sigma(:, z) + \sum_{x\in\mathcal{Y}} \alpha_x \sigma(:, x) \in \operatorname{span}\{\{\sigma(:, x)\}_{x\in\mathcal{Y}}, \sigma(:, z)\}$, for which $v_z(\Delta) = (0, \ldots, 0)$ and $\tilde{v}_z(\Delta^c) = v_z$ with some $v_z \in \mathbb{R}^{\Delta^c}$. Note that $2^N = \langle \sigma(:, z), \sigma(:, z)\rangle = \langle \tilde{v}_z, \sigma(:, z)\rangle = \langle v_z, \sigma(\Delta^c, z)\rangle$. For all $y \in \mathcal{Y}^c \setminus \{z\}$ we have that $\sigma(:, y) \perp \operatorname{span}\{\{\sigma(:, x)\}_{x\in\mathcal{Y}}, \sigma(:, z)\}$, and therefore $\sigma(\Delta^c, y) \perp v_z \quad \forall z \neq y, \ z, y \in \mathcal{Y}^c$. Summarizing, there exists a set of vectors $\{v_z\}_{z\in\mathcal{Y}^c}$ such that $\langle \sigma(\Delta^c, y), v_z\rangle = 2^N \delta_{y,z} \quad \forall y, z \in \mathcal{Y}^c$. Written as a matrix multiplication this is: $\left[v_{z_1}, \ldots, v_{z_{|\mathcal{Y}^c|}}\right]^\top \cdot \sigma(\Delta^c, \mathcal{Y}^c) = 2^N \operatorname{diag}(\mathbb{1})$. We have $|\mathcal{Y}^c| = |\Delta^c|$, such that $\sigma(\Delta^c, \mathcal{Y}^c)$ is square. From $\det(A \cdot B) = \det(A) \cdot \det(B)$, $\det \sigma(\Delta^c, \mathcal{Y}^c) \neq 0$ so that it has full rank.

Now consider an arbitrary $\mathcal{Y}$ with $|\mathcal{Y}| \leq |\Delta|$, (otherwise use $\mathcal{Y}' = \mathcal{Y}^c$ and $\Delta' = \Delta^c$). By Corollary 1.C.4, $\sigma(\Delta, \mathcal{X})$ has full rank $|\Delta|$, since $|\mathcal{X}| = 2^N > (1 - \frac{1}{|\Delta|})2^N$. There exists a set $\tilde{\mathcal{Y}}$ with $\mathcal{X} \supseteq \tilde{\mathcal{Y}} \supseteq \mathcal{Y}$, $|\tilde{\mathcal{Y}}| = |\Delta|$ and $\operatorname{rk} \sigma(\Delta, \tilde{\mathcal{Y}}) = |\Delta|$. By the first part of the proof, this is equivalent to $\operatorname{rk} \sigma(\Delta^c, \tilde{\mathcal{Y}}^c) = |\Delta^c|$. But this implies $\operatorname{rk} \sigma(\Delta^c, \mathcal{Y}^c) = |\Delta^c|$. The reverse direction is analogue.

Item (iii) For $\sigma(\Delta, \mathcal{Y}_\Delta)$: By item *(i)*, $\langle \sigma(2^\Lambda, x), \sigma(2^\Lambda, y)\rangle = 2^{|\Lambda|}\delta_{x(\Lambda),y(\Lambda)}$ for any $\Lambda \subseteq [N]$. If $\Delta = 2^\Lambda$, then the matrix $\sigma(\Delta, \mathcal{Y}_\Delta)$ has orthogonal rows. For the general case, denote by $\mathscr{F}$ the set maximal inclusion subsets of Δ. For any $x, y \in \mathcal{Y}_\Delta$ there exists a pair $\Lambda, \Lambda' \in \mathscr{F}$ with $\operatorname{supp}(x - \mathbb{1}) \subseteq \Lambda$ and $\operatorname{supp}(y - \mathbb{1}) \subseteq \Lambda'$. The pair x, y is either equal, or $\sigma(2^{\Lambda''}, x) \perp \sigma(2^{\Lambda''}, y)$ for $\Lambda'' = \Lambda$ or $\Lambda'' = \Lambda'$. Consider some vector $\sigma(\Delta, x)$ for which x satisfies $\operatorname{supp}(x - \mathbb{1}) \subseteq \Lambda \in \mathscr{F}$. We show that no linear combination of vectors $s := \sum_{y\in\mathcal{Y}_\Delta\setminus\{x\}} \alpha_y \sigma(\Delta, y)$ can be equal to $\sigma(\Delta, x)$: $s(\Delta) \neq \sigma(\Delta, x)$. For any $y \in \mathcal{Y}_\Delta \setminus \{x\}$ we have *(a)* $\sigma(2^\Lambda, y) \perp \sigma(2^\Lambda, x)$ or *(b)* $\sigma(2^\Lambda, y) = \sigma(2^\Lambda, x)$ and $\exists \Lambda' \in \mathscr{F}, \operatorname{supp}(y - \mathbb{1}) \in 2^{\Lambda'}$ s.t. $\sigma(2^{\Lambda'}, y) \perp \sigma(2^{\Lambda'}, x)$. If we had $s(\Delta) = \sigma(\Delta, x)$, then the coefficients α_y of all y fulfilling *(a)* had to be 0. For all coefficients y which do not fulfill *(a)*, we have that they fulfill *(b)*. Since $\sigma(2^{\Lambda'}, y) \perp \sigma(2^{\Lambda'}, x)$, again, if $s(\Delta) = \sigma(\Delta, x)$, then all the α_y to these y had to be 0. And so forth. This completes the proof for $\sigma(\Delta, \mathcal{Y}_\Delta)$. The claim for $\sigma(\Delta^c, \mathcal{Y}_\Delta^c)$ follows using item *(ii)*.

Item (iv) a) By definition, $\sigma(\lambda, x) = \prod_{i\in\lambda} x_i = (-1)^{|\operatorname{supp}(\mathbb{1}-x)\cap\lambda|} \quad \forall \lambda \in \Delta$. On the other hand, $\sigma(\lambda, -x) = \prod_{i\in\lambda} -x_i = (-1)^{|\operatorname{supp}(\mathbb{1}+x)\cap\lambda|} = (-1)^{|\lambda|-|\operatorname{supp}(\mathbb{1}-x)\cap\lambda|} \quad \forall \lambda \in \Delta$. Therefore, $\sigma(\Delta, x) = \operatorname{diag}\left((-1)^{|\lambda_1|}, \ldots, (-1)^{|\lambda_{|\Delta|}|}\right) \cdot \sigma(\Delta, -x)$.

b) For an arbitrary $\xi \in \{\pm 1\}^{|\mathcal{Y}|}$ and Δ consisting of sets of equal cardinality, $\sigma(\Delta, \mathcal{Y}) = [\sigma(\Delta, \operatorname{supp}(\xi + \mathbb{1})), \sigma(\Delta, \operatorname{supp}(\xi - \mathbb{1}))]$. On the other hand

$$\sigma(\Delta, \mathcal{Y}^\xi) = [\sigma(\Delta, \operatorname{supp}(\xi + \mathbb{1})), \operatorname{diag}\left((-1)^{|\lambda_1|}, \ldots, (-1)^{|\lambda_{|\Delta|}|}\right) \cdot \sigma(\Delta, \operatorname{supp}(\xi - \mathbb{1}))] \,,$$

where the diagonal matrix can be replaced by $+1$ or -1.

Item (v) For any $x \in \mathcal{X}$, $\lambda \in 2^{[n]}$ and $y \in \mathcal{X}$

$$\sigma(\lambda, x * y) = (-1)^{|(\operatorname{supp}(\mathbb{1}-x)\triangle\operatorname{supp}(\mathbb{1}-y))\cap\lambda|} = (-1)^{|\operatorname{supp}(\mathbb{1}-y)\cap\lambda|}(-1)^{|\operatorname{supp}(\mathbb{1}-x)\cap\lambda|},$$

and thus $\sigma(\Delta, \mathcal{Y}^{(y)}) = \operatorname{diag}(\sigma(\Delta, y)) \cdot \sigma(\Delta, \mathcal{Y})$. The diagonal matrix is regular. □

Proof of Proposition 1.C.2. Note that $\mathcal{Y}_{\Delta_k}^{(y)} = B(k,y) \subseteq \mathcal{X}$. Any two balls of equal radius $k < N$ and different center are different.

For the second item: If $|\mathcal{Y}|$ is odd, then $\mathcal{Y} \neq \mathcal{X}$. Assume $y * \mathcal{Y} = \mathcal{Y}$. Then $\mathcal{Y} = \{x_1, \ldots, x_{\lfloor |\mathcal{Y}|/2 \rfloor}, y*x_1, \ldots, y*x_{\lfloor |\mathcal{Y}|/2 \rfloor}, x\}$, and $\mathcal{Y} = y*\mathcal{Y} = \{x_1, \ldots, x_{\lfloor |\mathcal{Y}|/2 \rfloor}, y*x_1, \ldots, y*x_{\lfloor |\mathcal{Y}|/2 \rfloor}, y*x\}$. Hence $y = \mathbb{1}$.

For the third item: Let $\tilde{\mathcal{Y}} := \mathcal{Y} \setminus \{x\}$ for some $x \in \mathcal{Y}$, such that $|\tilde{\mathcal{Y}}|$ is odd and all $\tilde{\mathcal{Y}}^{(y)}$, $y \in \mathcal{X}$ are different. There are at most $|\mathcal{Y}|$ ways to choose $\tilde{\mathcal{Y}}$ from $\mathcal{Y}$, and hence at least $\frac{2^N}{|\mathcal{Y}|}$ of the $\mathcal{Y}^{(y)}$, $y \in \mathcal{X}$ must be different. □

Proof of Lemma 1.C.3. Let $S^+ = \{j \colon v_j > 0\}$ and $s^+ = |S^+|$. Without loss of generality let $|\alpha_1| = \max_i\{|\alpha_i|\}$. We have then:

$$k\alpha_1^2 n \geq \sum_{i=1}^k \alpha_i^2 n = \sum_{i=1}^k \langle \alpha_i v^i, \alpha_i v^i \rangle = \langle \sum_{i=1}^k \alpha_i v^i, \sum_{i=1}^k \alpha_i v^i \rangle = \langle v, v \rangle$$
$$= \sum_{j=1}^n |v_j|^2 > \sum_{j \in S^+} |v_j|^2 = \frac{1}{s^+} \left(\sum_{j \in S^+} 1 \right) \left(\sum_{j \in S^+} |v_j|^2 \right) \geq \frac{1}{s^+} \left(\sum_{j \in S^+} |v_j| \right)^2 .$$

On the other hand we have the following:

$$2 \sum_{j \in S^+} |v_j| \overset{\dagger}{=} \sum_{j=1}^n |v_j| \geq \sum_{j=1}^n v_j v_j^1 = \sum_{j=1}^n \sum_{i=1}^k \alpha_i v_j^i v_j^1$$
$$= \sum_{i=1}^k \alpha_i \sum_{j=1}^n v_j^i v_j^1 = \sum_{i=1}^k \alpha_i \langle v^i, v^1 \rangle = \alpha_1 \langle v^1, v^1 \rangle = \alpha_1 n.$$

Inserting the last expression into the first equation yields $s^+ > \frac{n}{4k}$. For † we used $\langle (1, \ldots, 1), v \rangle = 0$ implies $\sum_{j=1}^n |v_j^+| = \sum_{j=1}^n |v_j^-|$. □

Hadamard models

In the remainder of this section $\mathcal{E}_\Delta$ denotes an exponential family on a set $\mathcal{X}$ of cardinality $|\mathcal{X}| = 2^N$ with and a sufficient statistics of the form

$$A = ((-1)^{|\mathrm{supp}(x) \cap \lambda|})_{\lambda \in \Delta, x \in \{0,1\}^N} . \tag{1.27}$$

We consider an arbitrary family $\Delta \subseteq 2^{[N]}$ (not necessarily a simplicial complex on $[N]$), such that A is a submatrix of a Hadamard matrix, but does not necessarily describe a hierarchical model. We ask for the cardinality of S-sets in the case that Δ is only assumed to have a certain cardinality.

Proposition 1.C.5. *Any $p \in \overline{\mathcal{P}(\mathcal{X})}$ with $|\mathrm{supp}(p)| < \frac{2^N}{2|\Delta^c|}$ is contained in $\overline{\mathcal{E}_\Delta}$ and $\mathrm{cs}(\mathcal{E}_\Delta)$ is $\left(\frac{2^N}{2|\Delta^c|} - 1\right)$-neighborly.*

Proof. By Lemma 1.C.3, for any $\Delta \subseteq 2^{[N]}$ we have $|\mathrm{supp}(m^+)| \geq \frac{2^N}{2|\Delta^c|}$ $\forall m \in \ker \sigma(\Delta, \mathcal{X}) \backslash \emptyset$. The claim follows from Lemma 1.2.5. □

Example 1.C.6. Let $\Delta \subseteq 2^{[N]}$ have 2^N-2 elements. By Lemma 1.C.1 *(i)*, $\langle \sigma(\lambda, \mathcal{X}), \sigma(\lambda', \mathcal{X}) \rangle = 0$ for all distinct $\lambda, \lambda' \in 2^{[n]}$. It follows that $|\{x \in \mathcal{X} : \sigma(\lambda, x) = \sigma(\lambda', x)\}| = |\{x \in \mathcal{X} : \sigma(\lambda, x) \neq \sigma(\lambda', x)\}|$. Therefore, $m = \sigma(\lambda, \mathcal{X}) + \sigma(\lambda', \mathcal{X})$ has $|\mathrm{supp}(m)| = \frac{2^N}{2}$ $\forall \lambda \neq \lambda'$. Since all entries of this m different from 0 have the same absolute value, and $\langle m, \sigma(\emptyset, \mathcal{X}) \rangle = 0$, we have that $|\mathrm{supp}(m^+)| = \frac{|\mathrm{supp}(m)|}{2}$.

If $|\Delta|$ is large, the convex support has few vertices with respect to its dimension. In this case there must exist large simplex faces of the convex support.

Proposition 1.C.7. *Let $\Delta \subseteq 2^{[n]}$ and $|\Delta| > \frac{2^N}{2}$. The convex support $\mathrm{cs}(\mathcal{E}_\Delta) = \mathrm{conv}\{\sigma(\Delta, x)\}_{x \in \mathcal{X}}$ contains at least $(\frac{2^N}{2|\Delta|-2^N})$ simplex faces of dimension $(2|\Delta| - 2^N - 1)$ and at least 2^N simplex faces of dimension $(2|\Delta| - 2^N - 2)$. Hence there are at least $(\frac{2^N}{2|\Delta|-2^N})$ S-sets of cardinality $(2|\Delta| - 2^N)$ and 2^N S-sets of cardinality $(2|\Delta| - 2^N - 1)$.*

Proof. The marginal polytope to $\Delta \subseteq 2^{[n]}$ is a $|\Delta|-1$-dim polytope with 2^N vertices. Kalai [67] showed the following result for general convex polytopes [67, Lemma 2.3]: *Every d-dim polytope with $d+b$ vertices contains a $(d-b+1)$-dimensional face which is a simplex.* This implies the existence of an S-set of cardinality $(2|\Delta| - 2^N - 1)$. Since this is odd, the claim follows from Lemma 1.2.10 and Lemma 1.C.2. □

Theorem 1.C.8. *Let $k \geq \frac{N}{2}$, $\Delta_k := \{\lambda \in 2^{[N]} : |\lambda| \leq k\}$, and $\Delta \subsetneq 2^{[N]}$ with $|\Delta| \geq |\Delta_k|$. For any $\kappa < 1$ there is an $N_0 = N_0(\kappa)$ such that for all $N \geq N_0$ there are 2^N different S-sets $\{\mathcal{Y}^{(y)} \subseteq \mathcal{X}\}_{y \in \mathcal{X}}$ of $\overline{\mathcal{E}_\Delta}$ with $|\mathcal{Y}^{(y)}| \geq 2^{\kappa N}$.*

Proof. We use Proposition 1.C.7. We need to show that for every κ there exists an $N(\kappa)$ with

$$2|\Delta| - 2^N \geq 2^{\kappa N}, \quad \forall \kappa < 1, \forall N \geq N_0(\kappa). \tag{1.28}$$

Note that $|\Delta_k| = \sum_{i=0}^{k} \binom{N}{i}$ and $|\Delta_k^c| = \sum_{i=k+1}^{N} \binom{N}{i} = \sum_{i=0}^{N-(k+1)} \binom{N}{i}$. Consider the worst case, where $|\Delta| = |\Delta_k|$. Then we have that the relation from eq. (1.28) is satisfied for k s.t. $2\sum_{i=0}^{k} \binom{N}{i} \geq 2^N + 2^{\kappa N}$. This equation is equivalent to $2^N \geq 2^{\kappa N} + 2\sum_{i=0}^{N-(k+1)} \binom{N}{i}$. The last term is at most $2^N - 2\binom{N}{\lfloor \frac{N}{2} \rfloor}$, $\forall k \geq N/2$, which is decreasing in N, given that $k \geq \frac{N}{2}$. Furthermore, for any fixed $\kappa < 1$ and N large enough, $\frac{2^{\kappa N}}{2^N}$ is arbitrarily close to zero. □

Proposition 1.A.3 characterizes S-sets of exponential families. The next proposition gives an alternative characterization of facial sets which are S-set:

Proposition 1.C.9. *Let $\Delta \subseteq 2^{[N]}$ and $\mathcal{Y} \subseteq \mathcal{X} = \{0,1\}^N$. Then $\mathrm{rk}\, \sigma(\Delta^c, \mathcal{Y}^c) = |\Delta^c|$, if*

(i) $\quad |\mathcal{Y}| < \dfrac{2^N}{|\Delta^c|}$, *or*

(ii) $\quad \Delta$ *is a simplicial complex, and $\mathcal{Y} \subseteq \mathcal{Y}_\Delta^{(y)}$ for some $y \in \mathcal{X}$.*

Hence if $\mathcal{Y}$ is facial and satisfies (i) or (ii), then $\mathcal{Y}$ is an S-set.

Proof. (i) By Corollary 1.C.4 (see Appendix 1.C), $\mathrm{rk}\, \sigma(\Delta^c, \mathcal{Y}^c) = |\Delta^c|$ whenever $|\mathcal{Y}^c| > (1 - 1/|\Delta^c|)2^N$,. *(ii)* Follows from Lemma (ii) and Lemma (v), since $\mathcal{Y} \subseteq \mathcal{Y}_\Delta$ implies $\mathcal{Y}^c \supseteq \mathcal{Y}_\Delta^c$. See also Lemma (iv). □

Example 1.C.10. For any $y \in \mathcal{X}$ the set $\mathcal{Y}_{\Delta_k}^{(y)}$ is the Hamming ball with radius k centered at y. The S-sets of the independence model are the intersection of the family of faces of the N-cube and the family of all sets of radius smaller or equal to the interaction order, $k = 1$, i.e., the sets contained in $\mathcal{Y}_{\Delta_1}^{(y)}$ for some $y \in \mathcal{X}$.

2 Convex Subsets, Secants, Geodesics and Convex Hulls

In this chapter we develop techniques to compute the smallest natural number m for which $\mathrm{Mixt}^m(\mathcal{E}) = \mathrm{conv}(\mathcal{E})$ for a general exponential family $\mathcal{E}$. This chapter contains a variety of individual results. The implications of some results to the main subject of this thesis are not fully elaborated at this moment and should be understood as a basis for future research. Hierarchical models (treated in Section 1.3) contain all point measures in their closures (when $\cup_{\lambda\in\Delta}\lambda = [n]$). They are always contained in the convex hull of their boundaries and allow a packing of their state spaces in terms of S-sets. In general, however, the convex hull of an exponential family is not the convex hull of its boundary and there does not exist a packing of their state spaces in terms of S-sets. In some cases the minimal S-set packing does not correspond to the smallest mixture that can represent any distribution (see Proposition 1.2.8). *How can we assess the Carathéodory number in these cases?* In Chapter 1 we investigated the support sets that can possibly arise from the mixtures. This is an approach to mixture models using the mixtures based at the boundary of the probability simplex. S-sets and the corresponding convex subsets in the closure of exponential families proved very helpful. *What can we say about convex subsets of exponential families?*

In Section 2.1 we characterize convex α-families and study relations between S-sets and convex subsets of exponential families. In Section 2.2 we give a description of the intersection of m-geodesics and exponential families (i.e., a description of secant lines and intersection points). We relate these intersections to the support sets of distributions in the closure of exponential families. Furthermore, we examine the convex hull and limit points of α-geodesics. In Section 2.4 we compute the Carathéodory number of some classes of exponential families which are not contained in the convex hull of their boundaries.

2.1 Convex Exponential and α-Families

Convex Exponential Families

A convex family on $\mathcal{X}$ is a subset $\mathcal{M} \subseteq \mathcal{P}(\mathcal{X}) \subset \mathbb{R}^{\mathcal{X}}$ with $\lambda p + (1-\lambda)q \in \mathcal{M}$ for all $\lambda \in [0,1]$ for all $p, q \in \mathcal{M}$. A convex exponential family is an exponential family which is a convex family. A partition of $\mathcal{X}$ is a collection of disjoint subsets of $\mathcal{X}$, denoted blocks, whose union is $\mathcal{X}$. Given a probability distribution $R \in \mathcal{P}(\mathcal{X})$ and a subset $\mathcal{Y} \subseteq \mathcal{X}$, the conditional probability distribution R *conditioned to* $\mathcal{Y}$ is $R(\cdot|\mathcal{Y}) := \frac{\mathbb{1}_{\mathcal{Y}} R(\cdot)}{R(\mathcal{Y})} \in \mathcal{P}(\mathcal{Y})$. The affine hull of two points $P, Q \in \mathcal{P}$ is $\mathrm{aff}\{P,Q\} := \{(1-\lambda)P + \lambda Q \colon \lambda \in \mathbb{R}\}$. The following partition was introduced by F. Matúš and N. Ay in [79] and will be important in our considerations:

Definition 2.1.1. Given two probability distributions $P, Q \in \mathcal{P}(\mathcal{X})$, the partition $\varrho_{P,Q}$ has blocks defined as the equivalence classes of the relation $x \sim y$ iff $P(x)Q(y) = P(y)Q(x)$ for any $x, y \in \mathcal{X}$. In other words $P(\cdot|\mathcal{Y}) = Q(\cdot|\mathcal{Y})$ holds for every block $\mathcal{Y} \in \varrho_{P,Q}$.

Proposition 2.1.2. (F. Matúš and N. Ay [79]). *Any convex exponential family supported by $\mathcal{X}$ has the following form:*

$$\mathcal{S}_{R,\varrho} := \Big\{ \sum_{\mathcal{Y}\in\varrho} \pi_{\mathcal{Y}} R(\cdot|\mathcal{Y}) \colon \pi_{\mathcal{Y}} > 0, \sum_{\mathcal{Y}\in\varrho} \pi_{\mathcal{Y}} = 1 \Big\} ,$$

where ϱ is a partition of $\mathcal{X}$ and $R \in \mathcal{P}(\mathcal{X})$.

$\mathcal{S}_{R,\varrho}$ is a convex family, because it is the relative interior of $\mathrm{conv}\{R(\cdot|\mathcal{Y})\}_{\mathcal{Y}\in\varrho}$. It is an exponential family with sufficient statistics $\{\mathbb{1}_{\mathcal{Y}} \colon \mathcal{Y} \in \varrho\}$ and reference measure R. The convex support $\mathrm{cs}(\mathcal{S}_{R,\varrho})$ is a $|\varrho - 1|$-dimensional simplex. The support sets of elements from $\overline{\mathcal{S}_{R,\varrho}}$ are $\mathcal{F}(\mathcal{S}_{R,\varrho}) = \{\cup_{\mathcal{Y}\in\tilde{\varrho}}\mathcal{Y} \colon \tilde{\varrho} \subseteq \varrho\}$ and depend only on the partition ϱ, but not on the strictly positive reference measure R. For any $\mathcal{Y} \in \varrho$ and $x \in \mathcal{Y}$ the moment map maps $\overline{\mathcal{P}}(\mathcal{Y})$ onto the vertex A_x. For any $\nu, \tilde{\nu} \in \mathcal{P}(\mathcal{X})$, $\mathcal{E}_{\nu,V}$ is a convex exponential family iff $\mathcal{E}_{\tilde{\nu},V}$ is a convex exponential family iff $\mathcal{F}(\mathcal{E}_{\nu,V}) = \mathcal{F}(\mathcal{E}_{\tilde{\nu},V}) = \{\cup_{\mathcal{Y}\in\varrho'}\mathcal{Y} \colon \varrho' \subseteq \varrho\}$ for some partition ϱ of $\mathcal{X}$.

The model $\mathcal{S}_{P,\varrho_{P,Q}}$ contains P and Q. This results from the choice $\pi_{\mathcal{Y}} = P(\mathcal{Y})$ (resp. $\pi_{\mathcal{Y}} = Q(\mathcal{Y})$) in the definition of $\mathcal{S}_{R,\varrho}$, which yields $P = \sum_{\mathcal{Y}} P(\mathcal{Y})P(\cdot|\mathcal{Y})$ (resp. $Q = \sum_{\mathcal{Y}} Q(\mathcal{Y})P(\cdot|\mathcal{Y})$). In [79] it is shown that $\mathcal{S}_{P,\varrho_{P,Q}}$ is the smallest convex exponential family containing $P, Q \in \mathcal{P}$. Based on a similar analysis it is possible to show that an exponential family which contains the convex set $\mathrm{conv}\{P, Q\}$ already contains $\mathcal{S}_{P,\varrho_{P,Q}}$. The following lemma resulted from personal discussions with J. Rauh:

Lemma 2.1.3. *Let $\mathcal{E}$ be an exponential family on $\mathcal{X}$ with sufficient statistics matrix A. Let $P, Q \in \mathcal{P}$. If $\mathcal{E} \supseteq \mathrm{conv}\{P, Q\}$, then $\mathcal{E} \supseteq \mathcal{S}_{P,\varrho_{P,Q}}$. Furthermore, $\mathcal{E}$ contains a d-dimensional convex set iff there exists a linear projection of the vectors $\{A_x\}_x$ onto the $(d+1)$ vertices of a d-dimensional simplex. If $\mathcal{E}$ contains a d-dimensional convex set, then $\mathrm{cs}(\mathcal{E})$ has a d-dimensional simplex face.*

We will further strengthen this result in Theorem 2.2.6, where we show that the assumption $\mathcal{E} \supseteq \mathrm{conv}\{P, Q\}$ can be relaxed to $|\mathcal{E} \cap \mathrm{aff}\{P, Q\}| \geq |\varrho_{P,Q}|$.

Proof of Lemma 2.1.3. (i) Let $\tilde{\mathcal{T}}$ be the *extended tangent space* of $\mathcal{E}$, i.e., the span of $\{\mathbb{1}, A_1, \ldots, A_d\}$, where $\{A_i\}$ are a sufficient statistics for $\mathcal{E}$. We assume that $P+\lambda(Q-P)$ is contained in $\mathcal{E}$ for every $\lambda \in [0, 1]$. Note that any element of $\mathcal{E}$ is a reference measure for $\mathcal{E}$. Therefore, $\log \frac{P+\lambda(Q-P)}{P}$ is contained in $\tilde{\mathcal{T}}$ for every $\lambda \in [0, 1]$. The difference of two vectors in $\tilde{\mathcal{T}}$ is also in $\tilde{\mathcal{T}}$, and hence, the n-th derivative of $\log \frac{P+\lambda(Q-P)}{P}$ for all $n > 0$. Let $N = |\varrho_{P,Q}|$. For any block $\mathcal{Y}$ of $\varrho_{P,Q}$ choose one $x_{\mathcal{Y}} \in \mathcal{Y}$. Define a matrix $M \in \mathbb{R}^{N\times N}$ with entries $M_{\mathcal{Y},n} = \left(\frac{Q(x_{\mathcal{Y}})-P(x_{\mathcal{Y}})}{P(x_{\mathcal{Y}})}\right)^n$. This is a regular Vandermonde matrix, since $\frac{Q(x_{\mathcal{Y}})-P(x_{\mathcal{Y}})}{P(x_{\mathcal{Y}})} \neq \frac{Q(x_{\mathcal{Z}})-P(x_{\mathcal{Z}})}{P(x_{\mathcal{Z}})}$ for any two different blocks $\mathcal{Y}$ and $\mathcal{Z}$ in $\varrho_{P,Q}$. This implies that the dimension of $\tilde{\mathcal{T}}$ is at least N, such that the dimension of $\mathcal{E}$ is at least $N - 1$. The family $\mathcal{S}_{P_{\mathcal{Y}}:\mathcal{Y}\in\varrho_{P,Q}}$ is $(N-1)$-dimensional and contains the functions $f^n(x) = M_{\mathcal{Y},n}\ \forall x \in \mathcal{Y} \in \varrho$ for all $n \in [N]$ in its extended tangent space. Hence $\mathcal{E} \supseteq \mathcal{S}_{P,\varrho_{P,Q}}$.

(ii) By the first item, if $\mathcal{E}$ contains a d-dimensional convex set, then it contains a convex exponential family $\mathcal{S}$ of dimension at least d. On the other hand $\mathcal{S}$ is an exponential subfamily of $\mathcal{E}$ iff the sufficient statistics of $\mathcal{S}$ is the image of A by some linear map, and its reference measure is in $\mathcal{E}$. The columns of any sufficient statistics matrix of a d-dimensional convex exponential family are the vertices of a d-dimensional simplex.

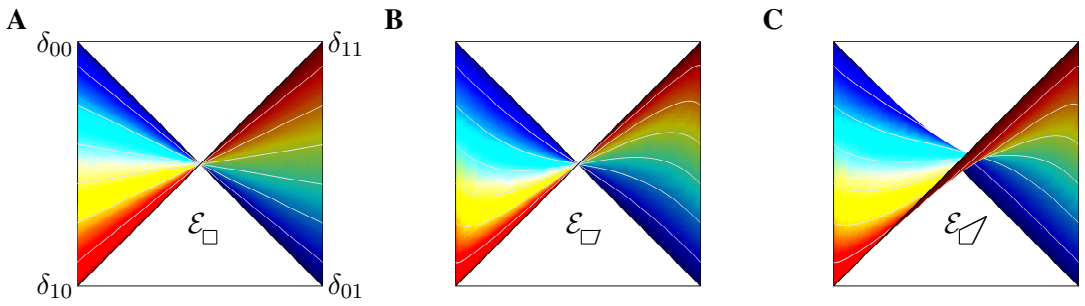

Figure 2.1: *This figure shows exponential families with uniform reference measure on* $\{0,1\}^2$*, projected along* $\operatorname{aff}\{\frac{1}{2}(\delta_{00}+\delta_{01}), \frac{1}{2}(\delta_{11}+\delta_{10})\}$ *(in the terminology of geometry of ruled surfaces this is the* line of striction *of* $\mathcal{E}^1_{2,\text{bin}}$*). The white lines are contours of* $p_{10}+p_{11}$*. The convex supports are: A parallelogram (A). In this case* $\mathcal{E}$ *is a doubly ruled surface. The depicted example is the binary independence model. A trapezoid which is not a parallelogram (B). In this case* $\mathcal{E}$ *is a (simply) ruled surface. A kite which is not a trapezoid (C). In this case* $\mathcal{E}$ *doesn't contain any non-trivial convex set, although its boundary consists of straight line segments.*

(iii) If there is a linear map of the columns of the sufficient statistics A onto the vertices of a d-dimensional simplex Δ, then the image of the d-dimensional faces of $\operatorname{cs}(\mathcal{E})$ covers Δ. Hence the image of one of these d-faces covers Δ and it must be a simplex. □

Remark 2.1.4. Two exponential families $\mathcal{E}$ and $\mathcal{E}'$ that have the same support sets $\mathcal{F}(\mathcal{E}) = \mathcal{F}(\mathcal{E}')$ do *not* necessarily satisfy: $\mathcal{E}$ contains a convex subfamily if $\mathcal{E}'$ does. See Figure 2.1.

Example 2.1.5. Let $\mathcal{E}$ be an exponential family on $\mathcal{X} = \{0,\dots,n-1\}$, $n \geq 5$, and let $\operatorname{cs}(\mathcal{E})$ be an n-gon with polyline $x \mapsto A_x$ (as in Example 1.2.7). There is no linear projection of an n-gon mapping the vertices onto the vertices of a d-simplex $d \geq 1$ (otherwise, 3 vertices of the n-gon would lie on a line). Hence $\mathcal{E}$ contains no convex sets of dimension larger than zero. The closure $\overline{\mathcal{E}}$ contains n one-dimensional convex sets given by $\operatorname{conv}\{\delta_x, \delta_{x+1}\} \mod n$.

By Example 2.1.5, a $(d+1)$-dimensional exponential family which contains a d-dimensional face of $\mathcal{P}$ in its closure, does not always contain a d-dimensional convex set. The following proposition explains this in more detail:

Proposition 2.1.6. *Let* $\mathcal{E}$ *be an exponential family on* $\mathcal{X}$*. If* $\mathcal{E}$ *contains a* d*-dim convex set,* $d \geq 1$*, then* $\mathcal{E}$ *has two facial sets which partition* $\mathcal{X}$*.*

More generally one can state: If $\overline{\mathcal{E}}$ contains $\operatorname{conv}\{Q, Q'\}$, $Q \neq Q'$, then there exist $\mathcal{Y}, \mathcal{Y}' \in \mathcal{F}(\mathcal{E})$ such that $\operatorname{supp}(Q) \cup \operatorname{supp}(Q') = \mathcal{Y} \dot\cup \mathcal{Y}'$.

Proof. If $\mathcal{E}$ contains a d-dim convex set, $d \geq 1$, then $\mathcal{E}$ contains $\operatorname{conv}\{Q, Q'\}$ for some $Q \neq Q'$. Hence, $\mathcal{E}$ contains $\mathcal{S}_{Q,\varrho_{Q,Q'}}$. Clearly $|\varrho_{Q,Q'}| \geq 2$, and hence $\operatorname{cs}(\mathcal{S}_{P,\varrho})$ is a simplex of dimension at least one. There exist two disjoint faces of $\operatorname{cs}(\mathcal{S}_{P,\varrho})$ covering all vertices. This corresponds to a partition of $\mathcal{X}$ into two disjoint facial sets of $\mathcal{S}_{P,\varrho}$. The claim follows from $\mathcal{F}(\mathcal{S}_\varrho) \subseteq \mathcal{F}(\mathcal{E})$. The second item follows from the first item and the observation that, for any $\mathcal{Z} \in \mathcal{F}(\mathcal{E})$, the intersection $\overline{\mathcal{E}} \cap \mathcal{P}(\mathcal{Z})$ is an exponential family on $\mathcal{Z}$. □

Any element of an exponential family is a reference measure of that exponential family, i.e., $\mathcal{E}_{\nu,A} = \mathcal{E}_{P,A}$ for any $P \in \mathcal{E}_{\nu,A}$. If $\mathcal{E}$ contains $\mathcal{S}_{R,\varrho}$, then $\mathcal{E}$ contains $\mathcal{S}_{Q,\varrho}$ for any $Q \in \mathcal{E}$. This can be used to construct convex foliations:

Proposition 2.1.7. *Let $\mathcal{E}$ be an exponential family with sufficient statistics A and extended tangent space $\tilde{\mathcal{T}} := \operatorname{span}\{\mathbb{1}, A_1, \ldots, A_d\}$. If $\mathcal{E}$ contains the convex family $\mathcal{S}_{P,\varrho}$, then $\mathcal{E} = \cup_{Q\in\mathcal{G}}\mathcal{S}_{Q,\varrho}$, where $\mathcal{G}$ is an exponential family with extended tangent space $\{\mathbb{1}_{\mathcal{Y}} : \mathcal{Y} \in \varrho\}^{\perp} \cap \tilde{\mathcal{T}} + \mathbb{R}\mathbb{1}$.*

Example 2.1.8. (Convex subsets of independence models).

(i) Consider the independence model of 2 binary variables, $\mathcal{E}^1_{2,\text{bin}}$. The entries of a vector $P = (P_{00}, P_{01}, P_{10}, P_{11})$ in this model are given by $P_{11} = p_1p_2$, $P_{10} = p_1(1-p_2)$, etc. The relation $\frac{P_{11}}{P_{10}} = \frac{Q_{11}}{Q_{10}}$ is equivalent to $p_2 = q_2$. Therefore, if $p_2 = q_2$, then $\varrho_{P,Q} = \{\{(10),(11)\},\{(00),(01)\}\}$, and $\mathcal{S}_{P,\varrho_{P,Q}} = \operatorname{aff}\{P,Q\} \cap \mathcal{P}$. The e-geodesic connecting P and Q is equal to the convex hull $\operatorname{conv}\{P,Q\}$. Similar arguments apply for the case $p_1 = q_1$. From this we see that $\mathcal{E}^1_{2,\text{bin}}$ contains two different straight lines through each of its points, i.e., it is a *doubly ruled surface*. An alternative way to see this is by considering the following unitary map (see also [120, page 130]):

$$f: \quad \mathcal{E}^1_{2,\text{bin}} \to \mathbb{R}^4; \quad f(P) = \frac{1}{2}\begin{pmatrix} 1 & 1 & 1 & 1 \\ -1 & -1 & 1 & 1 \\ -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \cdot \begin{pmatrix} P_{00} \\ P_{01} \\ P_{10} \\ P_{11} \end{pmatrix} .$$

The image of this map can be written as $f(\mathcal{E}^1_{2,\text{bin}}) = \{(\frac{1}{2}, r, s, 2rs) : -\frac{1}{2} < r, s < \frac{1}{2}\}$ using that $P^2_{00} - P^2_{01} - P^2_{10} + P^2_{11} = P_{00} - P_{01} - P_{10} + P_{11}$. This corresponds to a portion of the hyperbolic paraboloid $\{(x,y,z) : z = \frac{x^2}{2} - \frac{y^2}{2}\}$ with $r = \frac{1}{2}(x-y)$ and $s = \frac{1}{2}(x+y)$, and is known to be a doubly ruled surface (see e.g., [39]). Interestingly, the mean curvature of this surface is $H(x,y) = \frac{-x^2+y^2}{2(1+x^2+y^2)^{3/2}}$ and vanishes only for $x^2 = y^2$, implying that $\mathcal{E}^1_{2,\text{bin}}$ is not a minimal surface, (the mean curvature of minimal surfaces vanishes everywhere, see [90]).

(ii) Consider $\mathcal{E}^1_{n,\text{bin}}$. As explained in Section 1.1, page 16, the functions $\{A_i(x) = (-1)^{x_i} : i \in [n]\}$ are a sufficient statistics of this model and the convex support $\operatorname{conv}\{A_x\}$ is an n-dim cube $C_n = \operatorname{conv}\{\pm 1\}^n$. The cube has no simplex faces of dimension more than one, and hence the binary independence model contains no convex sets of dimension larger than one. On the other hand, for each $i \in [n]$, $\operatorname{diag}(0,\ldots,0,\underset{i}{1},0,\ldots,0)$ is a linear projection of C_n onto the interval $[-1,1]$ mapping $[x_i = 0]$ onto the point $\{1\}$, and $[x_i = 1]$ onto the point $\{-1\}$. Hence, for any $i \in [n]$, $\mathcal{E}^1_{n,\text{bin}}$ contains one-dimensional convex exponential subfamilies with facial sets $[x_i = 0]$ and $[x_i = 1]$. Therefore $\mathcal{E}^1_{n,\text{bin}}$ contains n different straight lines through each of its points ($\mathcal{E}^1_{n,\text{bin}}$ is n-ruled). For each $i \in [n]$ we have a convex foliation $\mathcal{E}^1_{n,\text{bin}} = \cup_{Q\in\mathcal{G}}\mathcal{S}_{Q,\varrho}$, where $\varrho = \{[x_i = 0],[x_i = 1]\}$ and $\mathcal{G}$ is the exponential family with sufficient statistics $\{A_j(x) = (-1)^{x_j} : j \in [n] \setminus \{i\}\}$.

(iii) The independence model $\mathcal{E}^1$ on $\bigtimes_{i\in[n]} \mathcal{X}_i$ is a $\big(\sum_i(|\mathcal{X}_i|-1)\big)$-dim model. Its convex support is a product of $(|\mathcal{X}_i|-1)$-dim simplices, $\operatorname{cs}(\mathcal{E}^1) = \bigtimes_{i\in[n]} \Delta^{|\mathcal{X}_i|-1}$, and can be projected onto $0 \times \cdots \times 0 \times \underset{i}{\Delta^{|\mathcal{X}_i|-1}} \times 0 \times \cdots \times 0$. Hence this model contains n different convex sets of dimensions $(|\mathcal{X}_i|-1)$, $i \in [n]$ through each of its points.

Interestingly, for each $j \in [n]$, there exists a linear projection M which maps $\mathcal{E}^1$ onto a set of dimension smaller than $\min\{\dim \mathcal{E}^1, \operatorname{rank} M - 1\}$. Marginalizing out the n-th variable

corresponds to the linear map $\mathcal{P}(\bigtimes_{i\in[n]}\mathcal{X}_i) \to \mathcal{P}(\bigtimes_{i\in[n-1]}\mathcal{X}_i)$; $p \mapsto M\cdot p$, where M is a matrix with rows $\{\mathbb{1}_{\{x\colon x_i=y_i\ \forall i\neq n\}}\}_{(y_1,\dots,y_{n-1})}$. It maps $p^1\cdots p^n$ onto $p^1\cdots p^{n-1}$ and maps $\mathcal{E}^1(\mathcal{X}_1\times\cdots\times\mathcal{X}_n)$ onto $\mathcal{E}^1(\mathcal{X}_1\times\cdots\times\mathcal{X}_{n-1})$.

Convex α-Families

The notion of α-families provides a class of models which interpolates between exponential and mixture families:

Definition 2.1.9. For any $(n+1)$ probability distributions $\{p_0,\dots,p_n\}\subset\mathcal{P}$, the *mixture family* ($\mathrm{m}$-family) $\mathcal{M}_{\{p_0,\dots,p_n\}}$ consists of all probability distributions of the following form:

$$p(x,\theta) = \sum_{i=1}^{n}\theta^i p_i(x) + (1-\sum_{i=1}^{n}\theta^i)p_0(x)\ , \quad \theta^i\in\mathbb{R}\ . \tag{2.1}$$

This is the secant space through $p_0,\dots,p_n$. An important class of mixture families results from the restriction $\theta^i>0$ and $\sum\theta^i<1$. Any m-family is convex, and the smallest m-family containing the convex hull of any P and Q is just $\{tP+(1-t)Q\colon t\in[0,1]\}=\mathrm{conv}\{P,Q\}$. A mixture model is a union of mixture families with basis points from a previously specified model. This should be compared to an exponential family: $\{p_\theta(x)=\exp\left\{C(x)+\sum_{i=1}^{d}\theta^iA_i(x)-\psi(\theta)\right\}\ \forall x\in\mathcal{X}\colon\theta\in\mathbb{R}^d\}$, which is characterized by the affine space $C+\mathrm{span}\{A_i\}_i\subseteq\mathbb{R}^{\mathcal{X}}$ modulo the constant functions. In this notation $\nu(x)=\exp(C(x))$ is a strictly positive reference measure and $\log(\psi)$ is the normalization constant.

An e-geodesic is a connected portion of a one-dimensional exponential family: $\{p_t(x)=\exp(C(x)+tA(x)-\psi(t))\ \forall x\in\mathcal{X}\colon t\in(a,b)\subset\mathbb{R}\}$, where $A\in\mathbb{R}^{\mathcal{X}}$ and $\mathbb{1}$ are linearly independent. If $p_{t=a}=P$ and $p_{t=b}=Q$ for two points $P,Q\in\mathcal{P}$, we say that $\{p_t\colon t\in[a,b]\}$ is an e-geodesic between P and Q. The natural parameters of an exponential family $\mathcal{E}$ form an e-affine coordinate system. Any e-geodesic between two points of $\mathcal{E}$ is contained in $\mathcal{E}$ and is given by a straight line segment in the natural parameter space of $\mathcal{E}$. We say that exponential families are e-geodesically convex. An m-geodesic is a standard straight line segment in $\mathcal{P}$. In general, given an *affine connection* on a manifold, the one-dimensional autoparallel submanifolds are called geodesics. See [10] for details on affine coordinates and interesting relations between mixture families and exponential families.

For each $\alpha\in\mathbb{R}$, Amari [10] defines

$$l^{(\alpha)}(x;\xi) := L^{(\alpha)}(p(x;\xi)) \quad\text{and}\quad L^{(\alpha)}(p) := \begin{cases}\frac{2}{1-\alpha}p^{\frac{1-\alpha}{2}} & \text{for } \alpha\neq 1\\ \log p & \text{for } \alpha=1\end{cases}\ . \tag{2.2}$$

Definition 2.1.10. An n-dimensional manifold $\mathcal{M}\subseteq\mathcal{P}(\mathcal{X})$ is an *α-family* iff there exist functions $\{C,A_0,\dots,A_n\}\subset\mathbb{R}^{\mathcal{X}}$ such that the *denormalization* $\widetilde{\mathcal{M}} := \{\tau p\colon p\in\mathcal{M},\tau>0\}$ satisfies

$$L^{(\alpha)}(\tau p(x;\xi)) = C(x)+\sum_{i=0}^{n}\theta^i(\xi,\tau)A_i(x) \quad \forall\tau p\in\widetilde{\mathcal{M}}\ , \tag{2.3}$$

where $\theta(\xi,\tau)$ is a one-to-one mapping of $\{(\xi_1,\dots,\xi_n,\tau)\}$, the natural parameters of $\widetilde{\mathcal{M}}$.

Within this framework, an exponential family is a (1)-family and a mixture family is a (-1)-family. The probability simplex $\mathcal{P}$ is an α-family for every $\alpha \in \mathbb{R}$. The α-geodesic between two points $p(x;\xi_0), p(x;\xi_1) \in \mathcal{P}$ is defined through the relation $l^{(\alpha)}(x;\xi_t) = (t-1)l^{(\alpha)}(x;\xi_0) + tl^{(\alpha)}(x;\xi_1)$.

For convex α-families we have the following:

Lemma 2.1.11. *Any convex exponential family $\mathcal{S}_{R,\varrho}$ is an α-family for every $\alpha \in \mathbb{R}$. For $\alpha \neq -1$ the smallest α-family containing the convex hull of a pair of distributions P and Q supported by $\mathcal{X}$ is $\mathcal{S}_{P,\varrho_{P,Q}}$. Hence any convex α-family is of the form $\mathcal{S}_{R,\varrho}$.*

Proof. $L^{(\alpha)}(\sum_i \pi_i R_i) = \sum_i \theta^i A_i$, choosing $A_i(x) = L^{(\alpha)}(R_i)$ for $x \in \mathcal{X}_i$ and $A_i(x) = 0$ else, and $\theta^i = \pi_i^{\frac{1-\alpha}{2}}$ for $\alpha \neq 1$ and $\theta_i = \log \pi_i$ for $\alpha = 1$.

The case $\alpha = 1$ is Lemma 2.1.3. For any $\alpha \neq 1$ we can write the elements of an α-family $\mathcal{G}$ as $p(x;\xi) = \left(\sum_{\lambda=0}^n \theta^\lambda(\xi) A_\lambda(x)\right)^{\frac{2}{1-\alpha}}$ [10, pg.49]. Let $\mathcal{T} = \operatorname{span} A_\lambda$, which corresponds to the extended tangent space of the α-family $\mathcal{G}$. If $\mathcal{G}$ contains the convex hull of P and Q, then $\mathcal{T}$ contains the vectors $(P + t(Q-P))^{\frac{1-\alpha}{2}}$ for all $t \in [0,1]$. The differences of vectors from that set must also be contained in $\mathcal{T}$ and thus the derivatives with respect to t. For $\alpha \neq -1$, the n-th derivative at $t = 0$ is $k_n P^{\frac{1-\alpha}{2}} \left(\frac{Q-P}{P}\right)^n$, where $k_n = \frac{1-\alpha}{2} \cdots (\frac{1-\alpha}{2} - (n-1))$. We have that $\frac{Q(x)-P(x)}{P(x)} = \frac{Q(y)-P(y)}{P(y)}$ iff $P(x)Q(y) = P(y)Q(x)$, and hence, the matrix $V_{i,n} = (\frac{Q(x_i)-P(x_i)}{P(x_i)})^n$ has full rank N. This implies that $\dim \mathcal{G} \geq N-1$, while the model $\mathcal{S}_{R,\varrho_{P,Q}}$ is an α-family of dimension $(N-1)$ and which contains P and Q. □

Corollary 2.1.12. *Let $\gamma_\alpha(P,Q)$ be the α-geodesic connecting two full support distributions P and Q. Then $\gamma_1(P,Q) = \gamma_{-1}(P,Q)$ iff $\gamma_\alpha(P,Q) = \gamma_{-1}(P,Q)\ \forall \alpha$ iff $\gamma_1(P,Q) = \{\exp(C(x) + tA(x) - \psi_t) \colon t \in \mathbb{R}\}$, where $A = \mathbb{1}_{\mathcal{Y}} - \mathbb{1}_{\mathcal{X}\setminus\mathcal{Y}}$.*

Example 2.1.13. Binary hierarchical models have a sufficient statistics which consists of rows of a Hadamard matrix (see Chapter 1). For $\mathcal{E}^k_{n,\text{bin}}$ these are $\{A_\lambda(x) = \prod_{i\in\lambda}(-1)^{x_i} \colon \lambda \subset [n], |\lambda| \leq k\}$. For every one of these A_λ there is a one-dimensional convex α-family contained in $\mathcal{E}^k_{n,\text{bin}}$.

Example 2.1.14. Consider the binomial model for n independent Bernoulli trials and probability p for success, $\text{Bin}_{n,p}(x) = \binom{n}{x} p^x (1-p)^{n-x}$, $x \in \{0,1,\ldots,n\}$. For $p \in (0,1)$, this model is a one-dimensional exponential family with sufficient statistics $A(x) = x$, natural parameter $\theta = \log p/(1-p)$ and reference measure $\nu(x) = \binom{n}{x}$. Any two points P and Q on this curve satisfy $\frac{P}{Q} \propto \exp((\theta_P - \theta_Q)x)$. Therefore, $|\varrho_{P,Q}| = n+1$ (unless $P = Q$). If an α-family ($\alpha \neq -1$) contains the convex hull of any two different points on $\text{Bin}_{n,p}$, then it equals the full probability simplex $\{\sum_{x\in\mathcal{X}} \pi_{\{x\}} P(x|\{x\}) \colon \pi\} = \mathcal{P}(\mathcal{X})$. The result extends to the case where one of the points lies on the boundary, i.e., $p_Q = 0$, or $p_Q = 1$. When $p_Q = 0$ and $p_P = 1$, we have $Q = \delta_0$, $P = \delta_n$, and there are many e-geodesics connecting the two points.

2.2 Secants of Exponential Families

We explore secants of exponential families, and more generally, intersections of different α-families. A secant line of some set $\mathcal{M}$ is a line which intersects $\mathcal{M}$ at two points; in other words, an m-geodesic that intersects $\mathcal{M}$ at two (or more) points. We use the following notation: Given

an exponential family $\mathcal{E}$ supported by $\mathcal{X}$ and a facial set $\mathcal{Y} \in \mathcal{F}(\mathcal{E})$, the *truncation* of $\mathcal{E}$ to $\mathcal{Y}$ is $\mathcal{E}_{\mathcal{Y}} := \{\frac{\mathbb{1}_{\mathcal{Y}} p}{\sum_{y\in\mathcal{Y}} p} : p \in \mathcal{E}\}$.

The following is a known fact about intersections of exponential families:

Example 2.2.1.

(i) Let $\mathcal{E}$ and $\mathcal{E}'$ be two exponential families on $\mathcal{X}$ with extended tangent spaces $\mathcal{T}$ and $\mathcal{T}'$ respectively, ($\mathcal{T} = \operatorname{span}\{\mathbb{1}, A_1, \dots, A_d\}$, where A is a sufficient statistics of $\mathcal{E}$). Then $\mathcal{E} \cap \mathcal{E}' = \tilde{\mathcal{E}}$ is an exponential family with extended tangent space $\tilde{\mathcal{T}} = (\mathcal{T}^{\perp} + \mathcal{T}'^{\perp})^{\perp} + \mathbb{R}\mathbb{1}$.

(ii) The closure of an exponential family $\overline{\mathcal{E}}$ is a disjoint union of exponential families $\mathcal{E}_{\mathcal{Y}} \subseteq \mathcal{P}(\mathcal{Y})$ defined on the facial sets $\mathcal{Y} \in \mathcal{F}(\mathcal{E})$. More precisely, $\overline{\mathcal{E}} = \cup_{\mathcal{Y}\in\mathcal{F}(\mathcal{E})}\mathcal{E}_{\mathcal{Y}}$, where the family $\mathcal{E}_{\mathcal{Y}}$ is the *truncation* of $\mathcal{E}$ to the set $\mathcal{Y}$ given by $\mathcal{E}_{\mathcal{Y}} := \{p \in \mathcal{P}(\mathcal{Y}) : p \propto q|_{\mathcal{Y}}, q \in \mathcal{E}\}$. Hence, the intersection of closures of exponential families satisfies $\overline{\mathcal{E}} \cap \overline{\mathcal{E}}' = \cup_{\mathcal{Y}\in\mathcal{F}(\mathcal{E})\cap\mathcal{F}(\mathcal{E}')}\tilde{\mathcal{E}}_{\mathcal{Y}}$, where $\tilde{\mathcal{E}}_{\mathcal{Y}}$ is the intersection $\mathcal{E}_{\mathcal{Y}} \cap \mathcal{E}'_{\mathcal{Y}}$ as described in the first item.

Example 2.2.2. The three exponential families shown in Figure 2.4 intersect at the one-dimensional (and planar) exponential family with sufficient statistics $(0, 1, -1, 0)$.

We are interested in the following question:

Question 2.2.3. What kinds of lines intersect exponential families and at how many points?

Intersection of m-Geodesics and Exponential Families

If two one-dimensional exponential families intersect at two points, then they are identical. More generally, if two α-geodesics intersect at more than one point in $\mathcal{P}$, then the two geodesics are identical. Now, at how many points can a one-dimensional α-family intersect another α'-family?

Lemma 2.2.4.

- *If an* e*-geodesic and an* m*-geodesic intersect at more than two points in $\mathcal{P}$, then they intersect at all of their points.*
- *More generally, all α-geodesics through a pair of points $P, Q \in \mathcal{P}$ either intersect only at P and Q, or are all equal, equal to the* m*-geodesic.*

Proof. Assume that an e-geodesic γ_{e} intersects an m-geodesic at three full support probability distributions P, Q and $P + \lambda(Q - P)$, for some $\lambda \in (0, 1)$. Then $\gamma_{\mathrm{e}} = P \exp(\theta \log \frac{Q}{P} - \psi_\theta)$, and for some reals θ and K the following holds: $\theta \log \frac{Q}{P} + K = \log(\frac{P+\lambda(Q-P)}{P})$. This yields the following equations on K and θ:

$$e^K z_i^\theta = (1 - \lambda) + \lambda z_i \quad \forall i \in \{1, \dots, |\mathcal{X}|\} \,, \tag{2.4}$$

where $(z_i)_i = z := \frac{Q}{P} \in \mathbb{R}^{\mathcal{X}}_{>0}$ is a vector with positive entries, and $\lambda \in (0, 1)$. For $\lambda \in \{0, 1\}$ the equation is solvable for any z. Now, as a function of $z_i > 0$ and for any K, $e^K z_i^\theta$ is linear if $\theta = 1$, strictly convex if $\theta > 1$ and strictly concave if $\theta < 1$. Hence, for any $\lambda \notin \{0, 1\}$, the curve $e^K z_i^\theta$ intersects $(1-\lambda)+\lambda z_i$ at most at two values of z_i, independently of how we choose K and θ. This is, the equations 2.4 are only solvable if the function $\frac{Q}{P}$ takes not more than two values. This is equivalent to $P(\cdot|\mathcal{Y}) = Q(\cdot|\mathcal{Y})$ and $P(\cdot|\mathcal{X} \setminus \mathcal{Y}) = Q(\cdot|\mathcal{X} \setminus \mathcal{Y})$ for some $\mathcal{Y} \subset \mathcal{X}$, and $|\varrho_{P,Q}| = 2$. But in this case we know that $\mathcal{S}_{P,\varrho_{P,Q}}$ is a one-dimensional exponential family containing P and Q. In turn, $\gamma_{\mathrm{e}} = \mathcal{S}_{P,\varrho_{P,Q}}$ is convex, and hence equal to the m-geodesic. For $\alpha \notin \{1, -1\}$ the claim follows from $\nabla^{(\alpha)} = \frac{1+\alpha}{2}\nabla^{(1)} + \frac{1-\alpha}{2}\nabla^{(-1)}$, see [10]. □

Now we extend Lemma 2.2.4 in order to account for the intersection of m-geodesics and arbitrary exponential families. First we show the following lemma, which is used in the proof of Theorem 2.2.6:

Lemma 2.2.5. *Let $P, Q \in \mathcal{P}(\mathcal{X})$, $N := |\varrho_{P,Q}|$, $\varrho_{P,Q} = \cup_{k\in[N]}\mathcal{X}_k$ and $x_k \in \mathcal{X}_k$ for $k \in [N]$. Furthermore, let $d \in \mathbb{N}$ and $f_i = \log \frac{P+\lambda_i(Q-P)}{P}$ with $i \in \{1, \ldots, d\}$ and $0 < \lambda_1 < \ldots < \lambda_d = 1$. Then the vectors $f(x_k) = (f_i(x_k))_i$ are the N vertices of a cyclic d-polytope combinatorially equivalent to $C(N, d)$.*

Proof. Let $\varrho_{P,Q} = \{\mathcal{X}_k\}_{k=1}^N$. For each k we choose a representative of the respective block $x_k \in \mathcal{X}_k$. Consider the following *alternant matrix* (an $n \times m$ matrix is called alternant if its entries contain the evaluation of m functions at n points):

$$f(i,k) := \log \frac{P(x_k) + \lambda_i(Q(x_k) - P(x_k))}{P(x_k)} = \log(1 + \lambda_i y_k) \ , \quad \forall\, k \in [N], \forall\, i \in [d] \ , \tag{2.5}$$

where $y_k := \frac{Q(x_k)-P(x_k)}{P(x_k)} > -1$. We show that the determinants $|(f_i(y_k))_{i,k\in[n]}|$ never vanish for $0 < \lambda_1 < \cdots < \lambda_n \leq 1$ and $-1 < y_1 < \cdots < y_n < \infty$. This is the case if the number of zeros of $\sum_{i=0}^n a_i f_i$ in the interval $-1 < y < \infty$ is at most n. A stronger condition is that the functions f_i satisfy *Descartes' rule* on the interval $-1 < y < \infty$, i.e., if $a_1, \ldots, a_n$ are reals not all equal to zero, then the number of zeros of $\sum_{i=1}^n a_i f_i$ in $-1 < y < \infty$ is at most equal to the number of times that the sequence $a_1, \ldots, a_n$ changes sign (disregarding zeros). By [93, ex. 87 and 90], the Descartes' rule is equivalent to the following condition: Given integers $1 \leq r_1 < \cdots < r_l \leq n$, the *Wronskian* $W[f_{r_i}(y)]$ doesn't vanish in $-1 < y < \infty$, and two determinants with the same number of rows have the same sign. We have

$$W[f_{r_i}(y)] := \begin{vmatrix} f_{r_1}(y) & f'_{r_1}(y) & \ldots & f^{(l-1)}_{r_1}(y) \\ \vdots & \vdots & & \vdots \\ f_{r_l}(y) & f'_{r_l}(y) & \ldots & f^{(l-1)}_{r_l}(y) \end{vmatrix} = \begin{vmatrix} \log(1+\lambda_{r_1} y) & \frac{\lambda_{r_1}}{(1+\lambda_{r_1} y)} & \ldots & \frac{(-1)^{l-2}(l-2)!\lambda_{r_1}^{l-1}}{(1+\lambda_{r_1} y)^{l-1}} \\ \vdots & \vdots & & \vdots \\ \log(1+\lambda_{r_l} y) & \frac{\lambda_{r_l}}{(1+\lambda_{r_l} y)} & \ldots & \frac{(-1)^{l-2}(l-2)!\lambda_{r_l}^{l-1}}{(1+\lambda_{r_l} y)^{l-1}} \end{vmatrix} .$$

This is a kind of *Vandermonde determinant*, and doesn't vanish iff $\frac{\lambda_{r_i}}{1+\lambda_{r_i} y}$ are different from each other. The later is true, since $\frac{z}{1+zy}$ is strictly increasing in $z \in [0, 1]$ whenever $y > -1$. The sign condition also is satisfied. If we multiply every second row by -1, the sign of the determinant is always strictly positive. This implies that the functions $\tilde{f}_i(y) = (-1)^i \log(1 + \lambda_i y)$ build a *Tchebycheff system* on $-1 < y < \infty$, i.e., $|(\tilde{f}_i(y_k))_{i,k\in[n]}| > 0$ (see [69, Page 25]). In particular, the map $f\colon (-1, \infty) \to \mathbb{R}^d \ ; \ y \mapsto (f_i(y))_i$ is a *d-order curve* in the sense of Sturmfels [108], i.e., $\begin{vmatrix} 1 & 1 & \ldots & 1 \\ f(y_1) & f(y_2) & \ldots & f(y_{d+1}) \end{vmatrix}$ is always positive (or always negative) whenever $-1 < y_1 < \cdots < y_{d+1}$. As pointed out in [108] (making reference to [82, 28]), the convex hull of N distinct points on any d-order curve is a combinatorial cyclic polytope. □

Theorem 2.2.6. (Intersections of exponential families and m-geodesics). *Let $P, Q \in \mathcal{P}(\mathcal{X})$ with $N := |\varrho_{P,Q}|$ and consider the line $\mathcal{L} := \mathrm{aff}\{P, Q\}$. If $\mathcal{L}$ intersects an exponential family $\mathcal{E}$ on $\mathcal{X}$ at $(d+1)$ points, then $\mathcal{E}$ contains an exponential subfamily with convex support combinatorially equivalent to the cyclic polytope $C(N, \min\{d, N-1\})$. In particular we have the following:*

- *If $\mathcal{L}$ intersects $\mathcal{E}$ at a finite number of points, then this number is at most $\dim(\mathcal{E}) + 1$.*

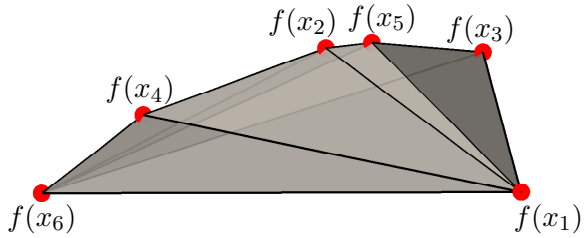

Figure 2.2: *The convex support of an exponential subfamily of any exponential family intersecting a straight line* $\mathrm{aff}\{P, Q\}$ *at the points* $(1-\lambda_i)P + \lambda_i Q$. *In this example* $\mathcal{X} = \{x_1, \ldots, x_6\}$, $P \propto (1,2,3,4,5,6)$, $Q \propto (3,1,3,1,3,1)$ *and* $(\lambda_i)_i = (0, 0.1, 0.9, 1)$. *This implies* $\varrho_{P,Q} = 6$. *An exponential family which intersects the mixture of* P *and* Q *at these points contains a subfamily with the convex support* $\mathrm{conv}\{f(x)\}_x$, *where* $f(x) = (f_i(x))_i$, $f_i(x) = \log \frac{P(x)+\lambda_i(Q(x)-P(x))}{P(x)}$. *Note that* $f_{\lambda=0}(x) \equiv 0$. *(The resulting, depicted polytope is much longer in one direction than in the others. This is not apparent in the chosen perspective, which highlights the faces of the polytope.)*

- *If* $\mathcal{L}$ *intersects* $\mathcal{E}$ *at* $|\varrho_{P,Q}|$ *points, then* $\mathcal{S}_{\nu,\varrho_{P,Q}} \subseteq \mathcal{E} \ \ \forall \nu \in \mathcal{E}$.
- *If* $\mathcal{L}$ *intersects* $\mathcal{E}$ *at more than one point, then* $\mathcal{E}$ *has* $|\varrho_{P,Q}|$ *facial sets which partition* $\mathcal{X}$.

Proof. Assume $\mathcal{L}$ intersects $\mathcal{E}$ at the following $(d+1)$ points: $P, P+\lambda_1(Q-P), \ldots, P+\lambda_d(Q-P) = Q$, where $0 = \lambda_0 < \lambda_1 < \cdots < \lambda_d = 1$. This implies that the tangent space of $\mathcal{E}$ contains the vectors $f_i = \log \frac{P+\lambda_i(Q-P)}{P}$ for $i \in \{0, \ldots, d\}$. The vectors f_i build the sufficient statistics of a subfamily $\mathcal{E}' \subseteq \mathcal{E}$. The properties of $\mathrm{cs}(\mathcal{E}')$ are shown in Lemma 2.2.5. By Lemma 2.2.5 the vectors f_i are linearly independent for $d+1 \leq |\varrho_{P,Q}|$, and hence $\dim(\mathcal{E}) \geq d$. If $d+1 = |\varrho_{P,Q}|$, then $\mathcal{E} \supseteq \mathcal{S}_{P,\varrho_{P,Q}}$, since $\dim(\mathcal{S}_{P,\varrho_{P,Q}}) = d$ and $\mathcal{S}_{P,\varrho_{P,Q}}$ contains all intersection points. The choice of the reference measure ν is arbitrary in $\mathcal{E}$. The third claim follows using $\mathcal{F}(\mathcal{E}') \subseteq \mathcal{F}(\mathcal{E})$ together with the fact that the columns $(f_i(x))_i$ are equal for all x in the same block of the partition $\varrho_{P,Q}$ and all columns are vertices of $\mathrm{cs}(\mathcal{E}')$. □

An exponential family which contains all $(k-1)$-dimensional faces of $\overline{\mathcal{P}}$ in its closure, $\overline{\mathcal{E}} \supset \cup_{\mathcal{Y}:1\leq|\mathcal{Y}|\leq k}\overline{\mathcal{P}}(\mathcal{Y})$, is a *k-Hamiltonian* or *k-neighborly* exponential family.

Remark 2.2.7.

(i) For a generic pair $P, Q \in \mathcal{P}(\mathcal{X})$, the partition $\varrho_{P,Q}$ has cardinality $|\mathcal{X}|$. This implies that an exponential family $\mathcal{E}$ intersecting a generic line at $d+1 < |\mathcal{X}|$ points contains an exponential subfamily whose convex support is combinatorially equivalent to the cyclic polytope $C(|\mathcal{X}|, d)$, and $\mathcal{E}$ is $\lfloor \frac{d}{2} \rfloor$-Hamiltonian. If $d+1 \geq |\mathcal{X}|$, then $\mathcal{E} = \mathcal{P}(\mathcal{X})$.

(ii) In the proof of Lemma 2.2.5 we in fact show that the functions $(-1)^i \log(1+\lambda_i y)$, $i \in [d]$, $0 < \lambda_1 < \ldots < \lambda_d = 1$ are a Tchebycheff system on $-1 < y < \infty$ (see [69, Page 25]).

Figure 2.2 illustrates Lemma 2.2.5 and Theorem 2.2.6. It shows the convex support of the smallest exponential family on a state space of cardinality six which intersects a generic straight line at four points.

Corollary 2.2.8. *If $\mathcal{E}$ has sufficient statistics matrix A and there is one x for which $A(x) \in \mathrm{ri}(\mathrm{cs}(\mathcal{E}))$, then any straight line intersects $\mathcal{E}$ at most at two points. Similarly, if there is one x for which $A(x) \in \mathrm{ri}(\mathrm{conv}\{A(y)\}_{y\in\mathcal{Y}})$ for some facial set $\mathcal{Y} \in \mathcal{F}(\mathcal{E})$, then any straight line intersects $\mathcal{E}_\mathcal{Y}$ at most at two points.*

Proof. This is immediate from the third item of Theorem 2.2.6, since $|\varrho_{P,Q}| \geq 2$ for any two distinct points on any straight line, while from the current assumption, the only facial set of $\mathcal{E}$ which contains x is $\mathcal{X}$. □

Corollary 2.2.9. *Let $\mathcal{E}$ be a two-dimensional exponential family which intersects a straight line at exactly three points. Then* $\mathrm{cs}(\mathcal{E})$ *is* monotone *in the following sense: There exist parallel supporting hyperplanes of two neighboring vertices of* $\mathrm{cs}(\mathcal{E})$.

Proof. Let $f(i,k) = \log \frac{P(x_k)+\lambda_i(Q(x_k)-P(x_k))}{P(x_k)}$ with $i \in \{1,\ldots,d\}$ and $0 < \lambda_1 < \ldots < \lambda_d = 1$, as in the Proof of Lemma 2.2.5. Here we have $d = 3$. For every $i \in [3]$ all entries of $(f(i,k))_{k\in[N]}$ are different. W.l.o.g. let $f(3,k) < f(3,l)$ for $k < l$. This entails $f(i,k) < f(i,l)$ for $k < l$ for all $i \neq 0$. We have the following:

$$\frac{f(i,k+1)-f(i,k)}{f(j,k+1)-f(j,k)} = \frac{\bar{f}'_{i,k}(y_{k+1}-y_k)}{\bar{f}'_{j,k}(y_{k+1}-y_k)} = \frac{\bar{f}'_{i,k}}{\bar{f}'_{j,k}}\,, \tag{2.6}$$

where $\bar{f}'_{i,k}$ denotes the mean slope of $y \mapsto \log(1+\lambda_i y)$ in the interval $[y_k, y_{k+1}]$. For $i < j$, the ratio of the slopes $\frac{f'_i(y)}{f'_j(y)} = \frac{\lambda_i+\lambda_i\lambda_j y}{\lambda_j+\lambda_i\lambda_j y}$ is strictly increasing in y. Hence, the expression of eq. (2.6) is strictly increasing in k, and the piecewise linear interpolation through the pairs $\{f(1,k), f(1,k+1)\}, 1 \leq k \leq N-1$ is a function. The claim is equivalent to the existence of a face G of $\mathrm{cs}(\mathcal{E})$ such that there exists a bijective orthogonal projection of $\partial\,\mathrm{cs}(\mathcal{E}) \setminus G$ into a straight line. This holds for $\mathrm{conv}\{f(x)\}$. Any regular linear map of $\{f(x)\}_x$ preserves this property, and hence the claim holds for the convex hull of the columns of any sufficient statistics of the exponential family. □

Consider a collection of probability distributions $P_0,\ldots,P_d \in \mathcal{P}$. The smallest exponential family containing all of them is $\mathcal{E}(P_0,\ldots,P_d) = \{\exp(\sum_i \theta_i \log(P_i) - \psi_\theta) \colon \sum_i \theta_i = 1\}$. This can be written as $\mathcal{E} = \{P_0 \exp(\sum_{i=1}^d \theta_i \log \frac{P_i}{P_0} - \psi_\theta) \colon \theta \in \mathbb{R}^d\}$. The functions $\{\log \frac{P_i}{P_0}\}_{i=1}^d$ are a sufficient statistics and P_0 a reference measure. If $\{\log \frac{P_i}{P_0}\}_{i=0}^d$ are linearly independent, then $\mathcal{E}$ is d-dimensional and the parametrization with θ is one-to-one. The exponential family depends only on the affine hull of $\{\log P_i\}_{i=0}^d$.

We can always define a d-dim exponential family which intersects a straight line at $d+1$ points, and in particular, a two-dimensional exponential family which intersects a straight line at three points. From Corollary 2.2.9 it follows that:

Example 2.2.10.

(i) An exponential family with convex support given by a regular n-gon, $n \geq 5$, intersects a straight line at most at two points.

(ii) A straight line $\mathcal{L}$ intersects $\mathcal{E}^1_{n,\mathrm{bin}}$ at zero, one, two, or at all points in $\mathcal{L} \cap \mathcal{P}$. See also Figure 2.1 A.

(iii) The families B and C shown in Figure 2.1 have monotone convex supports and there exist straight lines intersecting them at exactly three points.

2.3 α-Geodesics and α-Mixtures

Given two points P and Q in $\mathcal{P}$, denote $\gamma_\alpha(t)$ the α-geodesic between P and Q. The curve $\gamma_\alpha(t')$ for a fixed value t' and $\alpha \in [-1, 1]$ connects a point from the e-geodesic with a point from the m-geodesic. For $t' = 0$ or $t' = 1$, $\gamma_\alpha(t')$ is constant on α, equal to P and Q respectively. It is interesting to know whether the collection of curves $\gamma_\alpha(t)$ is contained in the convex hull of the e-geodesic. Furthermore, what can we say about the convex hull of α-geodesics, and the α-mixtures of exponential families?

Lemma 2.3.1. *Consider two distributions $P, Q \in \mathcal{P}$.*

- *For any α the convex hull of the α-geodesic between P and Q, is contained in $\mathcal{S}_{P,\varrho_{P,Q}}$ and hence it has a dimension* $\dim(\mathrm{conv}(\gamma_\alpha)) \leq |\varrho_{P,Q}| - 1$.
- *The smallest affine space containing γ_1 also contains $\mathcal{S}_{P,\varrho_{P,Q}}$. Hence, the convex hull of the* e*-geodesic has a dimension* $\dim(\mathrm{conv}(\gamma_\mathrm{e})) = |\varrho_{P,Q}| - 1$.

Proof of Lemma 2.3.1. From Lemma 2.1.11: The convex α-family $\mathcal{S}_{P,\varrho_{P,Q}}$ contains the α-geodesic between P and Q, because the later is the smallest α-family containing P and Q. In fact $\mathcal{S}_{P,\varrho_{P,Q}}$ contains a compact convex set which contains γ, and hence it also contains the convex hull of γ. Obviously $\dim \mathrm{conv}\gamma \leq \dim \mathcal{S}_{P,\varrho_{P,Q}} = |\varrho_{P,Q}| - 1$. For the lower bound: We have that $\varrho_{P,Q} = \{\mathcal{Y}\colon \frac{P(x)}{Q(x)} = \frac{P(y)}{Q(y)}\ \forall x, y \in \mathcal{Y}\}$. But this is equal to $\{\mathcal{Y}\colon \log \frac{P(x)}{Q(x)} = const.\, \forall x \in \mathcal{Y}\}$. □

An immediate consequence is that the α-geodesics between $P, Q \in \mathcal{S}$, $P \neq Q$ coincide for all α iff $|\varrho_{P,Q}| = 2$.

Consider some P and Q for which $|\varrho_{P,Q}| = 3$. In this case, all α-geodesics connecting P and Q are contained in a 2-dimensional affine space $\mathrm{aff}\{P(\cdot|A)\}_{A\in\varrho_{P,Q}}$. The e-geodesic has a sufficient statistics of the form $T = \mathbb{1}_{\mathcal{X}_1} - \mathbb{1}_{\mathcal{X}_2}$, $\mathcal{X}_1 \cap \mathcal{X}_2 = \emptyset$ and $\cup\mathcal{X}_i \subsetneq \mathcal{X}$.

Proposition 2.3.2. *Let P and Q be full support distributions with $|\varrho_{P,Q}| = 3$. Let γ_α denote the α-geodesic from P to Q. Then the following holds:*

$$\mathrm{conv}\gamma_\beta \subsetneq \mathrm{conv}\gamma_\alpha \quad \textit{for all} \quad -1 \leq \beta < \alpha \leq 1\ ,$$

and furthermore $\bigcup_{\beta\in[-1,\alpha]} \gamma_\beta = \mathrm{conv}\gamma_\alpha = \mathrm{Mixt}^2(\gamma_\alpha)$.

Proof. From Lemma 2.2.4: The set $\cup_{\alpha\in[-1,\beta]}\gamma_\alpha$ strictly increases with β. Furthermore, $[-1, \alpha] \ni \beta \mapsto \gamma_\beta$ is a homotopy of the two curves γ_α and γ_m with fixed points $\gamma_\beta(t = 0) \equiv P$ and $\gamma_\beta(t = 1) \equiv Q$. For any $\alpha \neq -1$, γ_α is a strictly convex plane curve, and the boundary of $\mathrm{conv}\gamma_\alpha$ is given by $\gamma_\alpha \cup \gamma_\mathrm{m}$. Any compact set S contains all extreme points of $\mathrm{conv}S$. From Proposition 2.3.1 we know that the convex hull of γ_α has dimension at most 2. Carathéodory's theorem says $\mathrm{Mixt}^3(\gamma_\alpha) = \mathrm{conv}(\gamma_\alpha)$. The mixtures of any 3 points on γ_α yield a triangle with vertices $\gamma_\alpha(t_1), \gamma_\alpha(t_2), \gamma_\alpha(t_3)$. The rays from $\gamma_\alpha(t_1)$ to the line interval $[\gamma_\alpha(t_2), \gamma_\alpha(t_3)]$ hit the curve γ_α, and hence any mixture of 3 points on γ_α can be written as a mixture of 2 points. □

If $|\varrho_{P,Q}| > 3$, then $\cup_{\beta\in[-1,\alpha)}\gamma_\beta \neq \mathrm{conv}(\gamma_\alpha)$, since $\dim(\cup\gamma_\beta) = 2$, while $\dim \mathrm{conv}(\gamma_\alpha) \geq 3$. Yet, there are interesting questions, such as: What is $\mathrm{conv}_\beta(\gamma_\alpha)$, i.e., the set that arises from adding β-geodesics between points indefinitely, starting from points in the set $\gamma\alpha$? Is it true that $\mathrm{conv}(\gamma_\beta) \subseteq \mathrm{conv}(\gamma_\alpha)$, $-1 \leq \beta < \alpha \leq 1$?

In Section 5.2 we will discuss two-dimensional 1-Hamiltonian exponential families (and m-dimensional k-Hamiltonian exponential families). Their convex supports have $|\mathcal{X}|$ vertices and they contain all point measures $\{\delta_x\}_{x\in\mathcal{X}}$ in their closures. The closure of any one-dimensional exponential family contains at most 2 point measures, and hence there don't exist one-dimensional 1-Hamiltonian exponential families for $|\mathcal{X}| > 2$. The following Lemma 2.3.3 states that for any $\mathcal{X}$, there exist one-dimensional exponential families which approximate all $\{\delta_x\}_{x\in\mathcal{X}}$ simultaneously to an arbitrary accuracy, and provides a one-dimensional counterpart to 1-Hamiltonian exponential families.

Lemma 2.3.3. (Exponential geodesics approaching all point measures). *For any finite $\mathcal{X}$ and any $\varepsilon > 0$ there exists an* e*-geodesic $\gamma = \{p_t \colon t \in \mathbb{R}\}$ for which $D(\delta_x\|\gamma) \leq \varepsilon\ \forall x \in \mathcal{X}$ and* $\mathrm{conv}(\gamma) \supseteq \mathcal{P}^\varepsilon := \{p \in \mathcal{P} \colon p(x) \geq \varepsilon\ \forall x \in \mathcal{X}\}$.

Figure 2.3 illustrates the result.

Proof of Lemma 2.3.3. Let $\mathcal{X} = \{1, \ldots, N\}$. We set $A(i) = i$ and $\nu(i) = \exp(\sum_{k=1}^{f_i} 10^{-k}K)$ for $i \in \mathcal{X}$, where

$$f_i = \begin{cases} i, & i \leq \lceil \frac{N}{2} \rceil \\ N-(i-1), & i > \lceil \frac{N}{2} \rceil \end{cases} .$$

The geodesic is $p_t(i) = \frac{\nu(i)e^{t\cdot i}}{\sum_j \nu(j)e^{t\cdot j}}$, $t \in \mathbb{R}$. The claim is that given any $\varepsilon > 0$ we always find a K such that for each $i \in \mathcal{X}$, there is a $t = t_i$ with $p_{t_i}(i) \geq (1-\varepsilon)$ and $p_{t_i}(j) < \varepsilon$ for all $j \neq i$. We show the following equivalent statement: Given any $\kappa \geq 1$, for all $i \in \mathcal{X}$ there exists a $t_i \in \mathbb{R}$ such that

$$\frac{p_{t_i}(i)}{p_{t_i}(j)} = \exp\big((i-j)t_i + (\sum_{k=1}^{f_i} 10^{-k} - \sum_{l=1}^{f_j} 10^{-l})K\big) > e^\kappa \quad \forall j \neq i\,. \tag{2.7}$$

It is sufficient to show that the exponent in the left hand side of the inequality can be made larger than zero for all $j \neq i$, since in this case we can multiply t and K by a constant in order to satisfy the inequality. Let $i \leq \lceil \frac{N}{2} \rceil$ (the proof for $i > \lceil \frac{N}{2} \rceil$ is analogue). There are four cases: *i)* $j_1 < i$, *ii)* $\lceil \frac{N}{2} \rceil \geq j_2 > i$, *iii)* $N-(i-1) > j_3 > \lceil \frac{N}{2} \rceil \geq i$, *iv)* $j_4 \geq N-(i-1)$. Inserting *iii)* into eq.(2.7) we get that t must be smaller than 0. In this case *iv)* is always satisfied. The remaining cases are satisfied if the following inequalities hold:

$$-\frac{\sum_{k=j_1+1}^{i} 10^{-k}}{(i-j_1)} < \frac{\sum_{k=i+1}^{j_2} 10^{-k}}{(i-j_2)}\ ,\ \frac{\sum_{k=i+1}^{N-(j_3-1)} 10^{-k}}{(i-j_3)}\,. \tag{2.8}$$

The left hand side is smaller than -10^{-i}, while the terms in the right hand side are larger than $10^{-j_2}\frac{j_2-(i+1)}{i-j_2} \geq -10^{-(i+1)}$, and larger than

$$10^{-N-(j_3-1)}\frac{(N-(j_3-1))-(i+1)}{i-j_3} \geq -10^{-(i+1)}\frac{\lceil \frac{N}{2} \rceil - 1 - i}{j_3-i} \geq -10^{-(i+1)}$$

respectively. This completes the proof. □

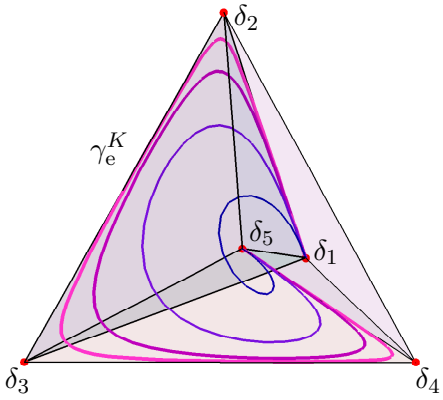

Figure 2.3: *This figure illustrates Lemma 2.3.3. It shows the Schlegel diagram of the four-dimensional simplex $\mathcal{P}(\mathcal{X} = \{1, \dots, 5\})$ and four elements from a one-parameter set of* e*-geodesics: For every $K \in \mathbb{R}$ γ_{e}^K is an* e*-geodesic with natural parameter $t \in \mathbb{R}$ and limit points $\gamma^K(t \to \pm\infty) = \{\delta_1, \delta_5\}$. The blue curve is for a small value of K and the magenta one is for a larger value of K. The distance from γ_{e}^K to any δ_x goes to zero as $K \to \infty$.*

Limit Points of α-Geodesics

There are many e-geodesics which hit the boundary of the probability simplex at the same points. Every one-dimensional exponential family with a sufficient statistics which attains a unique maximum value at x and a unique minimum value at y hits the boundary at δ_x and δ_y. On the other hand it is clear that different m-geodesics have different boundary points. Here we study the behavior of boundary points for general α-geodesics. We start with the collection of α-geodesics through a common pair of full support distributions P and Q.

Lemma 2.3.4. *Consider a pair of probability distributions $P, Q \in \mathcal{P}(\mathcal{X})$.*

- *The limit points of the* e*-geodesic $\gamma_{\mathrm{e}}(t)$ through P and Q are probability distributions with support* $\operatorname{argmax} \frac{P}{Q}$ *for $t \to \infty$ and* $\operatorname{argmin} \frac{P}{Q}$ *for $t \to -\infty$.*
- *For $\alpha \neq 1$. The α-geodesic $\gamma_\alpha(t)$ through P and Q hits the boundary of $\mathcal{P}$ at a probability distribution with support $\{x \in \mathcal{X} \colon \frac{P(x)}{Q(x)} \neq \min \frac{P}{Q}\}$ and at a probability distribution with support $\{x \in \mathcal{X} \colon \frac{P(x)}{Q(x)} \neq \max \frac{P}{Q}\}$.*

Remark 2.3.5. Given any pair of points, the support of boundary distributions of the e-geodesic through the two points is always contained in the support of the boundary distributions of α-geodesics through the same two points. Furthermore, for the two boundary points $\partial\gamma_\alpha^{(1)}$ and $\partial\gamma_\alpha^{(2)}$ the map $\alpha \mapsto \partial\gamma_\alpha^{(1/2)}$ is continuous.

Proof of Lemma 2.3.4. Let $\mathcal{X} = \{1, \dots, n\}$. We write $P = (p_i)$ and $Q = (q_i)$ and $\gamma_\alpha = \gamma = (\gamma_1, \dots, \gamma_{|\mathcal{X}|})$. First item: In the case $\alpha = 1$ we have $\log \gamma_i(t) = t \log p_i + (1-t) \log q_i - \psi(t)$, or equivalently $\gamma_i(t) = \frac{q_i \left(\frac{p_i}{q_i}\right)^t}{\sum_i q_i \left(\frac{p_i}{q_i}\right)^t}$. This means that $\gamma(t)$ hits the boundary of $\mathcal{P}$ at a probability distribution with support $\operatorname{argmax} \frac{p_i}{q_i}$ for $t \to \infty$ and at a probability distribution with support $\operatorname{argmin} \frac{p_i}{q_i}$ for $t \to -\infty$.

Second item: Consider some $\alpha \neq 1$. Then the α-geodesic through P and Q is a curve $\gamma(t) = (\gamma_i(t))_{i\in\mathcal{X}}$ which satisfies

$$(\gamma_i(t))^{\frac{1-\alpha}{2}} = C(t)\left(tp_i^{\frac{1-\alpha}{2}} + (1-t)q_i^{\frac{1-\alpha}{2}}\right) . \tag{2.9}$$

This curve hits the boundary of the probability simplex at two points. The first point is given by $\gamma(t_+)$, where t_+ is the smallest positive t for which $\gamma_i(t) = 0$ for some $i \in \mathcal{X}$. The second is $\gamma(t_-)$, where t_- is the largest negative t for which $\gamma_i(t) = 0$ for some i. The condition $\gamma_i(t) = 0$ can be written as $\bar{q}_i + t(\bar{p}_i - \bar{q}_i) = 0$, where $\bar{p}_i := p_i^{\frac{1-\alpha}{2}}$ and $\bar{q}_i := q_i^{\frac{1-\alpha}{2}}$. If $p_i = q_i \neq 0$, then $\gamma_i \neq 0$ for all t. For those i with $p_i \neq q_i$ we consider the following two sets: $I_+ := \{i\colon p_i > q_i\}$ and $I_- := \{i\colon p_i < q_i\}$. For $P \neq Q$ the two sets $I_\pm$ are not empty. We get the following:

$$t_\pm = \begin{cases} \min_{i\in I_-} -\bar{q}_i/(\bar{p}_i - \bar{q}_i) \\ \max_{i\in I_+} -\bar{q}_i/(\bar{p}_i - \bar{q}_i) \end{cases} . \tag{2.10}$$

Hence, at time t_+, the curve γ hits the boundary of $\mathcal{P}$ at a distribution which vanishes in $i = \operatorname{argmin} \frac{-\bar{q}_i}{\bar{p}_i - \bar{q}_i}\Big|_{I_-} = \operatorname{argmin} \frac{p_i}{q_i}\Big|_{I_-} = \operatorname{argmin} \frac{p_i}{q_i}$. Similarly, $\gamma(t_-)$ is a distribution which vanishes in $i = \operatorname{argmax} \frac{p_i}{q_i}$. □

Remark 2.3.6. By Lemma 2.3.4, unless $\alpha = 1$, given two points in the boundary of $\mathcal{P}$, there is not much freedom in the choice of an α-geodesic connecting the two points (straight lines, for example, are completely determined from the two points, in contrast to exponential families, which are only vaguely determined from their boundary points, unless the boundary points have complementary supports).

Proposition 2.3.7. *Consider an* e*-geodesic* $\gamma(t) = \nu\exp(tf - \psi_t)$, $\nu > 0$. *Let* $I_+ = \{x \in \mathcal{X}\colon f(x) = \max f\}$ *and* $I_- = \{x \in \mathcal{X}\colon f(x) = \min f\}$. *The limit points of* γ *have supports* $\operatorname{supp}(\gamma(-\infty)) = I_-$ *and* $\operatorname{supp}(\gamma(+\infty)) = I_+$. *Furthermore* $\gamma(\pm\infty)|_{I_\pm} \propto \nu|_{I_\pm}$.

Proof. This follows from Lemma 2.3.4. We provide an alternative proof: The support sets of the probability distributions in the boundary of the exponential family are I_- and I_+ because these sets are the index sets of all columns of the sufficient statistics matrix $f \in \mathbb{R}^{1\times\mathcal{X}}$ which lie in a proper face of the convex support $\operatorname{conv}\{f_x\}_{x\in\mathcal{X}} = [\min f, \max f]$ (see [48] and [96]). The second statement is clear from the fact that $\exp(tf(x))$ is constant on $\{x \in \mathcal{X}\colon f(x) = \max f\}$. □

Proposition 2.3.8. *Let* $P, Q \in \partial\mathcal{P}$, $P = \delta_x$ *and* $Q = \delta_y$ *for some* $x, y \in \mathcal{X}$, $x \neq y$. *The collection of all* e*-geodesics with limit points* P *and* Q *covers all* $\mathcal{P}$.

Proof. Let $P^0 = \exp(f_0 - \psi_0)$ and $P^1 = \exp(f_1 - \psi_1)$ be probability distributions in $\mathcal{P}$. The functions f_0 and f_1 are arbitrary (finite) elements of $\mathbb{R}^{\mathcal{X}}$. The e-geodesic connecting P^0 and P^1 is $P(t) = \exp((1-t)f_0 + tf_1 - \psi_t) = \exp(f_0)\exp(t(f_1 - f_0) - \psi_t)$. This is the one-dimensional exponential family with strictly positive reference measure $\nu = \exp(f_0)$ and sufficient statistics $(f_1 - f_0)$. The probability distribution $P(0) = \exp(f_0 - \psi_0) = \nu/\sum_x \nu(x)$ can be made arbitrary in $\mathcal{S}$ by choosing a suitable f_0. Furthermore, for any f_0 we may choose $f_1 = f_0 + \|f_0\|_1\mathbb{1}_x - \|f_0\|_1\mathbb{1}_y$, in which case the limit points of $P(t)$ are $P(+\infty) = \delta_x = P$ and $P(-\infty) = \delta_y = Q$. Hence, given any $\mu \in \mathcal{P}$, we find an e-geodesic with limit points P and Q, which contains $P(0) = \mu$. □

If the limit points have complementary supports, then the e-geodesic is unique:

Proposition 2.3.9. *If $P, Q \in \partial\mathcal{P}$ have complementary supports $\mathrm{supp}(P) = \mathcal{X} \setminus \mathrm{supp}(Q)$, then there is a unique* e*-geodesic with limit points P and Q: The* m*-geodesic between P and Q.*

Proof. Any e-geodesic with limit points P and Q must have a sufficient statistics A satisfying $\mathrm{argmax}(A) = \mathrm{supp}(P)$ and $\mathrm{argmin}(A) = \mathrm{supp}(Q)$. This implies that $A = \mathbb{1}_{\mathrm{supp}(P)}$ is a sufficient statistics. On the other hand $\nu = k_1 P + k_2 Q$ must be a reference measure. This is $\gamma_{\mathrm{e}}(t) = \frac{e^t k_1}{e^t k_1 + k_2} P + \frac{k_2}{e^t k_1 + k_2} Q$. □

In particular, Proposition 2.3.9 shows that a pair of boundary points of an exponential family are not necessarily connected by an exponential geodesic within the exponential family. In contrast, for pairs of strictly positive distributions, the exponential geodesic is contained in any exponential family containing them. This reflects the well known fact that the natural parametrization of an exponential family $\mathcal{E}$ doesn't extend to the boundary $\partial\mathcal{E}$. We discuss this further in Proposition 2.3.11.

In contrast to Proposition 2.3.8, given the limit points, the α-geodesics with $\alpha < 1$ are unique and the following holds:

Proposition 2.3.10. *Consider any $P, Q \in \partial\mathcal{P}(\mathcal{X})$, $P \neq Q$.*

- *For all $\alpha < 1$: There exists an α-geodesic in $\mathcal{P}$ connecting P and Q. This α-geodesic is contained in $\mathcal{P}(\mathrm{supp}(P) \cup \mathrm{supp}(Q))$.*
- *For $\alpha = 1$: There exists an* e*-geodesic connecting P and Q iff $\mathrm{supp}(P) \cap \mathrm{supp}(Q) = \emptyset$. It is unique iff $\mathrm{supp}(P) \cup \mathrm{supp}(Q) = \mathcal{X}$.*

Proof of Proposition 2.3.10. First item: There is no problem in writing the geodesic $\gamma(t)$ from equation 2.9 for $P, Q \in \partial\mathcal{P}$. On the other hand, from that equation we also see that if $P_i = Q_i = 0$ for some $i \in \mathcal{X}$, then $\gamma_i(t)\ \forall t$, and in consequence $\gamma \subseteq \mathcal{S}(\mathrm{supp}(P) \cup \mathrm{supp}(Q))$.

Second item: An e-geodesic $\gamma(t)$ can be given as a one-dimensional exponential family with arbitrary reference measure ν and an arbitrary sufficient statistics $f \in \mathbb{R}^{\mathcal{X}}$. Let $\mathcal{Y} = \mathrm{supp}(P)$ and $\mathcal{Z} = \mathrm{supp}(Q)$. For the *if* statement: We assume $\mathcal{Y} \cap \mathcal{Z} = \emptyset$. The choice $\nu|_{\mathcal{Y}} \propto P|_{\mathcal{Y}}$ and $\nu|_{\mathcal{Z}} \propto Q|_{\mathcal{Z}}$ together with $f = \mathbb{1}_{\mathcal{Y}} - \mathbb{1}_{\mathcal{Z}}$ yields that $\gamma(\infty) = \nu\mathbb{1}_{\mathcal{Y}} / \sum_{x\in\mathcal{X}} \mathbb{1}_{\mathcal{Y}}(x)\nu(x) = P$ and $\gamma(-\infty) = \nu\mathbb{1}_{\mathcal{Z}} / \sum_{x\in\mathcal{X}} \mathbb{1}_{\mathcal{Z}}(x)\nu(x) = Q$. For the *only if* statement: The support of a distribution belonging to the boundary of a one dimensional exponential family with sufficient statistics $f \in \mathbb{R}^{\mathcal{X}}$ is a subset of $\mathcal{X}$ consisting of those x for which $f(x)$ is a boundary point of the line segment $\mathrm{conv}\{f_x\}_{x\in\mathcal{X}} = [\min f, \max f]$, Proposition 2.3.7. Now, either f is a constant function, in which case the exponential family consists of only one point, or $\{x : f_x = \max f\} \cap \{x : f_x = \min f\} = \emptyset$, in other words, the support sets of the boundary distributions are disjoint. □

A natural question is if it is possible to define an exponential family as the exponential mixture of a number of its extreme points, or starting from a predetermined boundary. This is not possible (see, Proposition 2.3.8 and Proposition 2.3.9), at least not without further ado, since different exponential families can have the same boundary. Therefore, one usually considers the e-convex hull of a collection of full support probability distributions and the extension to the e-affine hulls, i.e., the span of a sufficient statistics.

An interesting question is which points in the boundary of $\mathcal{E}$ are limit points of e-geodesics parametrized by one-dimensional linear subspaces of the parameter space of $\mathcal{E}$. This is, by $\theta_r = r\vartheta$ for a fixed vector ϑ in the unit $(d-1)$-sphere S^{d-1} and $r \in \mathbb{R}_{\geq 0}$. It is also interesting to know how these limit points depend on the parametrization of $\mathcal{E}$.

Proposition 2.3.11. *Let $\mathcal{E}_{\nu,A}$ be the exponential family with reference measure ν and a sufficient statistics $\{A_i\}_{i=1}^d$. The set of limit points of* e*-geodesic with natural parameters of the form $\{r\vartheta \colon r \in \mathbb{R}\}$ for some $\vartheta \in S^{d-1}$ consists of the truncations of the reference measure to the sets $\mathcal{F}(\mathcal{E}) = \{\operatorname{argmax} f \colon f \in \operatorname{span}\{A_i\}\}$. For every $\mathcal{Y} \in \mathcal{F}(\mathcal{E})$ there is exactly one limit point $\nu\mathbb{1}_{\mathcal{Y}}/\sum_{x\in\mathcal{Y}}\nu$ with support $\mathcal{Y}$. On the other hand, every $p \in \partial\mathcal{E}$ is the limit point of an* e*-geodesic of the form given above by choosing an appropriate reference measure.*

Proof. The limit distributions for $r \to \infty$ and a fixed $\vartheta \in S^{d-1}$ are given by the truncation of ν to any set from $F_{\mathcal{E}} := \{\operatorname{argmax} f \colon f \in V\}$, Proposition 2.3.7. The truncation of ν to $\mathcal{Y} \in F_{\mathcal{E}}$ is defined as $\nu|_{\mathcal{Y}} := \nu\mathbb{1}_{\mathcal{Y}}/\sum_{y\in\mathcal{Y}}\nu(y)$. Furthermore, $\mathcal{Y} = \operatorname{argmax} f$ for some f in the span of $\{A_i,\}$ is equivalent to the existence of a vector $\theta \in \mathbb{R}^d$ such that $\sum_{i=1}^d \theta_i A_i(x) = 0\ \forall x \in \mathcal{Y}$ and $\sum_{i=1}^d \theta_i A_i(x) \leq -1\ \forall x \notin \mathcal{Y}$. This is precisely the definition of a facial set of $\mathcal{E}$. The claim follows from the fact that for a facial set $\mathcal{Y}$, $\overline{\mathcal{E}} \cap \mathcal{P}(\mathcal{Y})$ equals the truncation of $\mathcal{E}$ to $\mathcal{Y}$ [95, Theorem 2.29], and the fact that $\operatorname{cs}(\mathcal{E})$ and $\overline{\mathcal{E}}$ are homeomorphic. □

Example 2.3.12. Consider the product distributions of two binary variables. Let $x = (x_1, x_2) \in \{0,1\}^2 = \mathcal{X}$. Given three product distributions $P^i = (p_{i,1}p_{i,2}, p_{i,1}(1-p_{i,2}), (1-p_{i,1})p_{i,2}, (1-p_{i,1})(1-p_{i,2}))$, $i = 0,1,2$, we have $\mathcal{E}(P^0,P^1,P^2) := \{P^0 \exp(\theta_1 \log \frac{P^1}{P^0} + \theta_2 \log \frac{P^2}{P^0} - \psi)\}$. The expression in the exponent can be written as $\theta_1(x_1 \log \frac{p_{1,1}}{p_{0,1}} + (1-x_1)\log\frac{1-p_{1,1}}{1-p_{0,1}} + x_2 \log\frac{p_{1,2}}{p_{0,2}} + (1-x_2)\log\frac{1-p_{1,2}}{1-p_{0,2}})$ plus a similar expression for θ_2. We can write this in the following form:

$$(\theta_1,\theta_2)\cdot\begin{pmatrix} \log\frac{p_{1,1}p_{1,2}}{p_{0,1}p_{0,2}} & \log\frac{p_{1,1}(1-p_{1,2})}{p_{0,1}(1-p_{0,2})} & \log\frac{(1-p_{1,1})p_{1,2}}{(1-p_{0,1})p_{0,2}} & \log\frac{(1-p_{1,1})(1-p_{1,2})}{(1-p_{0,1})(1-p_{0,2})} \\ \log\frac{p_{2,1}p_{2,2}}{p_{0,1}p_{0,2}} & \log\frac{p_{2,1}(1-p_{2,2})}{p_{0,1}(1-p_{0,2})} & \log\frac{(1-p_{2,1})p_{2,2}}{(1-p_{0,1})p_{0,2}} & \log\frac{(1-p_{2,1})(1-p_{2,2})}{(1-p_{0,1})(1-p_{0,2})} \end{pmatrix}$$

The vectors $(1,1,1,1)$, $(1,1,-1,-1)$, $(1,-1,1,-1)$, $(-1,1,1,-1)$ build a basis of $\mathbb{R}^{\mathcal{X}}$ and $(-1,1,1,-1)$ is a kernel element of the matrix in the above-equation. Hence $\mathcal{E} = \{P_0 \exp(\theta' \cdot A) \colon \theta' \in \mathbb{R}^2\}$, where P^0 is any element of the exponential family and $A = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix}$. The facial sets are $\mathcal{X}$, $\{11,10\}$, $\{11,01\}$, $\{00,10\}$, $\{00,10\}$ and the individual elements of $\mathcal{X}$. For a fixed P^0 the e-geodesics with parameters given by a linear subspace are precisely those passing through P^0. The limit points are P^0 for the facial set $\mathcal{X}$, $(p_{0,1}p_{0,2}, p_{0,1}(1-p_{0,2}), 0, 0)/(p_{0,1}p_{0,2} + p_{0,1}(1-p_{0,2}))$ for the facial set $\{11,10\}$, and similar distributions for the other pairs. For the facial sets $\{x\} \in \mathcal{X}$ the limit distributions are δ_x, independently of P^0.

α-Mixtures of Exponential Families

We define α-geodesic mixtures (α-mixtures for short) recursively. In order to have a consistent notation for mixtures, we denote the α-geodesic connecting the points p and q by $\operatorname{Mixt}_\alpha(p,q)$. We define the n-th α-mixture of a set $\mathcal{G} \subseteq \mathcal{P}$ as

$$\operatorname{Mixt}_\alpha^n(\mathcal{G}) := \operatorname{Mixt}_\alpha(\mathcal{G}, \operatorname{Mixt}_\alpha^{n-1}(\mathcal{G}))\ , \tag{2.11}$$

with $\mathrm{Mixt}_\alpha(\mathcal{G}, \mathcal{G}') := \{\mathrm{Mixt}_\alpha(p, q) \colon p \in \mathcal{G}, q \in \mathcal{G}'\}$.

For $\alpha = -1$ we get the usual notion of mixtures. In Chapter 1 we showed that if the boundary of an exponential family $\mathcal{E}$ contains entire faces of $\mathcal{P}$, and $\kappa^s_{\mathcal{E}}$ is the smallest cardinality of a covering which contains all vertices of $\mathcal{P}$, then $n \geq \kappa^s_{\mathcal{E}} \;\Rightarrow\; \mathrm{Mixt}^n(\mathcal{E}) = \mathcal{P}$. It is possible to compute or estimate $\kappa^s_{\mathcal{E}}$ for various interesting exponential families. On the other hand, for an exponential family $\mathcal{E}$ it is $\mathrm{Mixt}^n_{\mathrm{e}}(\mathcal{E}) = \mathcal{E}$ for every n. What can we say about $\mathrm{Mixt}^n_\alpha(\mathcal{E})$?

Proposition 2.3.13. *For $\alpha < 1$, $\mathcal{P}(\mathcal{Y})$ is the smallest α-family which approaches every point measure δ_x, $x \in \mathcal{Y} \subseteq \mathcal{X}$. Hence $\mathcal{P}$ is the smallest α-family that approaches every point measure.*

Proof. By Proposition 2.3.10, an α-geodesic with $\alpha \neq 1$ and limit points δ_x and δ_y is contained in $\mathcal{P}(\{x, y\})$. Therefore, in fact this geodesic equals the interval $[\delta_x, \delta_y]$. Similarly, the smallest α-family which reaches this interval and a further point δ_z has closure $\mathrm{conv}\{\delta_y, \delta_y, \delta_z\}$. Induction yields the result. □

An α-family which contains all point measures in its closure, is equal to the full probability simplex, and we get a straight extension of Lemma 1.2.2 to the case of α-mixtures:

Corollary 2.3.14. *Let $\mathcal{E}$ be an exponential family. If $n \geq \kappa^s_{\mathcal{E}}$ then $\mathrm{Mixt}^n_\alpha(\mathcal{E}) = \mathcal{P}$ for any $\alpha < 1$.*

Proof. Let $\{\mathcal{Y}_i\}_i$ be S-sets for $\mathcal{E}$, i.e., the support sets of faces of $\overline{\mathcal{P}}$ which are contained in $\overline{\mathcal{E}}$. Assume that there are $\kappa^s_{\mathcal{E}}$ of them that cover $\mathcal{X}$. We can assume without loss of generality that they are disjoint. Proposition 2.3.13 implies that the α-mixture $\mathrm{Mixt}_\alpha(\overline{\mathcal{P}}(\mathcal{Y}_i), \overline{\mathcal{P}}(\mathcal{Y}_j))$ is equal to $\overline{\mathcal{P}}(\mathcal{Y}_i \cup \mathcal{Y}_j)$. Iteration of this yields $\mathrm{Mixt}^n_\alpha(\overline{\mathcal{E}}) = \overline{\mathcal{P}}$ for $n \geq \kappa^s_{\mathcal{E}}$. The result for strictly positive distributions follows the lines of the Proof of the Mixture Decompositions using S-sets Lemma from [86]. □

Remark 2.3.15. For α-mixtures of compact subsets of $\mathcal{E}$ we expect in general $\mathrm{Mixt}^n_\alpha(\mathcal{E}) \neq \mathrm{Mixt}^n_{\alpha'}(\mathcal{E})$ for $\alpha \neq \alpha'$.

2.4 Convex Hulls

Carathéodory Number of Some Exponential Families

If $\mathcal{E}$ is an exponential family with sufficient statistics A and there is an $x \in \mathcal{X}$ for which the column A_x is not a vertex of the convex support, i.e., $A_x \notin \mathrm{ex}(\mathrm{cs}(\mathcal{E}))$, then every mixture of elements of $\overline{\mathcal{E}}$ has support either not containing x, or strictly larger than x. It is interesting to know for which n the mixture model equals the convex hull. We use the results from the previous subsections to study two interesting situations. We use the following terminology:

Definition 2.4.1. Consider two sets $V \subset \mathbb{R}^d$ and $V' \subseteq \mathrm{conv}(V)$. The *Carathéodory number* of V' with respect to V, written $\mathrm{Car}_V(V')$, denotes the smallest natural number m for which any $p \in V'$ is the convex combination of at most m points in V. If $\mathrm{Car}_V(\mathrm{conv}(V)) = m$, we say that V has Carathéodory number m and write $\mathrm{Car}(V) = m$.

Consider a d-dimensional exponential family $\mathcal{E}$ for which and all but one columns of its sufficient statistics are vertices of $\mathrm{cs}(\mathcal{E})$. If the column A_{x_0} is contained in a k-dimensional face of

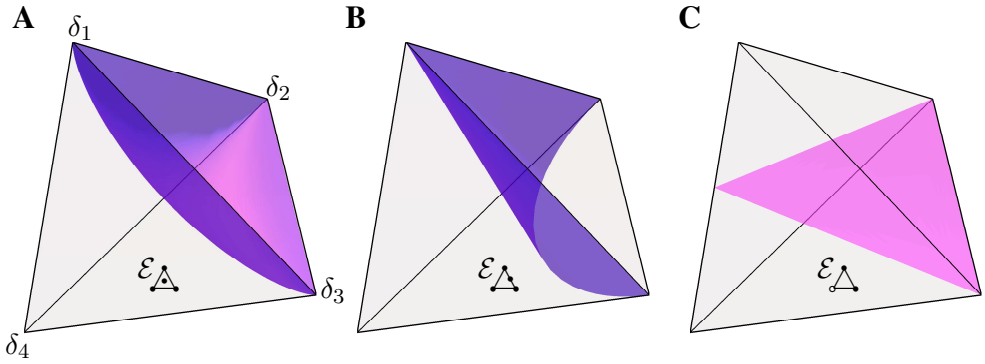

Figure 2.4: *This figure shows exponential families on $\mathcal{X} = \{1, 2, 3, 4\}$. They illustrate Proposition 2.4.2 and Theorem 2.4.3. For these families $p_\eta(4)$ is a concave function on the convex support. The small figures indicate the configuration of the column vectors of the sufficient statistics. Example 2.4.5 provides a more detailed discussion.*

$\mathrm{cs}(\mathcal{E})$, then $\mathrm{cs}(\mathcal{E})$ can be projected along this face (along basis vectors of the supporting hyperplane) onto a simplex of dimension $(d-k)$. This implies that $\mathcal{E}$ contains a convex subfamily $\mathcal{S}_{P,\varrho}$ attached to each $P \in \mathcal{E}$, see Lemma 2.1.3.

Proposition 2.4.2. *If $\mathcal{X} = \{x_0, x_1, \ldots, x_N\}$ and $\mathcal{E}$ is an exponential family with sufficient statistics A such that $\mathrm{cs}(\mathcal{E})$ is a d-pyramid with basis $\mathrm{conv}\{A_{x_i}\}_{i=1}^N$, then $\mathcal{E}$ contains the one-dimensional convex families with sufficient statistics $\mathbb{1}_{\{x_0\}}$. In fact $\mathcal{E}$ is $\mathrm{Mixt}(\mathcal{E}_{\mathcal{X}\setminus\{x_0\}}, \delta_{x_0})$ and $\mathrm{Car}(\mathcal{E}_{\mathcal{X}\setminus\{x_0\}}) = \mathrm{Car}(\mathcal{E})$, where $\mathcal{E}_{\mathcal{X}\setminus\{x_0\}}$ is the truncation of $\mathcal{E}$ to $\mathcal{X}\setminus\{x_0\}$.*

This result can be easily extended to the case where several columns of the sufficient statistics lie in the apex of the pyramid.

Proof. Use Proposition 2.3.9. □

Theorem 2.4.3. *Let $\mathcal{E}$ be an exponential family on $\mathcal{X}$, $|\mathcal{X}| = d+2$, with sufficient statistics A, for which $\mathrm{cs}(\mathcal{E})$ is a d-dimensional simplex. If $A_y \notin \mathrm{ex}(\mathrm{cs}(\mathcal{E}))$ for one $y \in \mathcal{X}$, the map $\mathrm{cs}(\mathcal{E}) \to [0,1]\,;\, \eta \mapsto p_\eta(y)$ is concave and $\mathrm{Car}(\mathcal{E}) = 2$. If $A_y \in \mathrm{ex}(\mathrm{cs}(\mathcal{E}))$ for all $y \in \mathcal{X}$, then $\mathcal{E}$ is a simplex and $\mathrm{Car}(\mathcal{E}) = 1$.*

Remark 2.4.4.

(i) The map $\mathbb{R}^d \to \mathbb{R}\,;\, \theta \mapsto p_\theta(y)$, $p_\theta = \exp(\theta \cdot A - \psi_\theta)$ is not concave.

(ii) In the case $|\mathcal{X}| \geq d+3$ (e.g., the binomial model Bin_n, $n \geq 3$), the map $\eta \mapsto p_\eta(y)$ is not necessarily convex, concave or monotone for a non-facial $\{y\}$.

Proof of Theorem 2.4.3. The assumption implies that there is a set $\mathcal{X}' := \mathcal{X}\setminus\{y\}$, $y \in \mathcal{X}$ such that $\{A_x\}_{x\in\mathcal{X}'}$ are vertices of $\mathrm{cs}(\mathcal{E})$, and A_y lies in the convex hull of $\{A_x\}_{x\in\mathcal{X}'}$. The polytope $\mathrm{cs}(\mathcal{E})$ is the disjoint union of the relative interiors of its faces, $\mathrm{cs}(\mathcal{E}) = \dot\cup_{G\in\mathcal{F}(\mathrm{cs}(\mathcal{E}))} \mathrm{ri}(G)$, where $\mathrm{ri}(\{v\}) = \{v\}$. Let G be the face in which A_y is contained, $A_y \in \mathrm{ri}(G)$. If $G = A_{x'}$, then $\mathcal{E}$ is a convex family with blocks $\{x', y\}$ plus the atoms of $\mathcal{X}\setminus\{x', y\}$. Consider now $G = \mathrm{cs}(\mathcal{E})$. $\mathcal{E}$ is a manifold of codimension one in $\mathcal{P}$ and $\partial\mathcal{E} \subset \partial\mathcal{P}$. The Jordan-Brouwer separation theorem yields that $\mathcal{E}$ divides $\mathcal{P}$ into two regions $\mathcal{P}^+$ and $\mathcal{P}^-$. Since A_y is in the

interior of $\mathrm{cs}(\mathcal{E})$, Theorem 2.2.6 yields that any line intersects $\mathcal{E}$ at most at two points. A line which intersects $\mathcal{P}(\mathcal{X}')$ at two points is contained in $\mathcal{P}(\mathcal{X}')$. W.l.o.g. $\overline{\mathcal{P}}^+ \supset \overline{\mathcal{P}}(\mathcal{X}')$. A regular line must intersect the boundary $\partial\mathcal{P}^+ = \mathcal{E} \cup \overline{\mathcal{P}}(\mathcal{X}')$ at an even number of points. Any line segment joining two points of $\mathcal{E}$ is contained in $\overline{\mathcal{P}}^+$, and in turn $\overline{\mathcal{P}}^+$ is a convex set. Now we show that $\mathrm{Mixt}^2(\mathcal{E}) \supset \mathcal{P}^+$. There are 2 S-sets covering $\mathcal{X}'$, and hence $\mathrm{Mixt}^2(\overline{\mathcal{E}}) \supseteq \overline{\mathcal{P}}(\mathcal{X}')$. The normal space of $\mathcal{E}$, given by $\ker A$, is one-dimensional. The union of the fibers fills the simplex $\uplus_{p\in\overline{\mathcal{E}}}\mathcal{N}_p = \overline{\mathcal{P}}$ and the union of the positive parts of the fibers fills $\mathcal{P}^+$. For any $p \in \mathcal{E}$ with $A \cdot p = \eta \in \mathrm{cs}(\mathcal{E})$ there are two points in $\partial\,\mathrm{cs}(\mathcal{E})$ with $\sum \pi_i\eta_i = \eta$ and hence $\sum \pi_i p_{\eta_i} = q$ with $A \cdot q = \eta$. Furthermore, $q \in \partial\mathcal{P}$, because $p_{\eta_i}(y) = 0$. This implies $\{q, p\} = \partial\mathcal{N}_p^+$. Since $\mathcal{N}_p$ is one dimensional, the curve $\gamma_{p_1,p_2}(t) := \sum \pi_i p_{(1-t)\eta_i+t\eta} \subseteq \mathcal{N}_p$ with end points $\{q, p\}$ in fact contains $\mathcal{N}_p^+$. The remaining cases of G follow from recursive application of Proposition 2.4.2. □

The Kullback-Leibler divergence from $P \in \overline{\mathcal{P}}$ to $\mathcal{E}$ is $\inf_{Q\in\mathcal{E}} D(P\|Q)$, where $D(P\|Q) := \sum_x P(x)\log\frac{P(x)}{Q(x)}$. S. Weis [120] discussed codimension one exponential families for which the centroid of $\mathcal{P}(\mathcal{X}\setminus\{x\})$ is a local Kullback-Leibler maximizer. The cardinality of the support of a local maximizer is at most $\dim(\mathcal{E}) + 1$, a bound which is attained in this case. J. Rauh [95] shows that any local maximizer of KL-divergence to an exponential family $\mathcal{E}(\mathcal{X})$ belongs to the class of *kernel distributions*[1] $K_\mathcal{E}$, and that $K_\mathcal{E} \cap \overline{\mathcal{P}}(\mathcal{Y})$ is a convex set which is not empty iff $\mathcal{Y} \subset \mathcal{X}$ is not facial. For any codimension one exponential family there exist exactly two local maximizers of KL-divergence. For $\mathcal{E}$ as in Theorem 2.4.3 we have

$$\max_{p\in\overline{\mathcal{P}}} D(p\|\mathrm{conv}(\mathcal{E})) = D(\delta_y\|\mathcal{E}) \ ,$$

and δ_y is one of two local maximizers of $D_\mathcal{E}$.

Example 2.4.5. We discuss the three families on $\mathcal{X} = \{1, 2, 3, 4\}$ depicted in Figure 2.4. They have codimension one in $\mathcal{P}$ and do not contain every δ_x in their closure. The dots in  ,  and  represent the configuration of column vectors of the sufficient statistics. The convex hull of these models does not fill the entire probability simplex. We are interested in $\mathrm{Car}(\mathcal{E})$. In the following examples A and B Theorem 2.4.3 yields $\mathrm{Car}(\mathcal{E}) = 2$, while in example C the family is convex. **A.** The sufficient statistics has rows $(1, 0, 0, \frac{1}{3})$ and $(0, 1, -1, 0)$. The vector $A_{x=4}$ lies is in the relative interior of $\mathrm{cs}(\mathcal{E})$. This exponential family can be thought of as a higher dimensional analogue to the Hardy-Weinberg exponential family on the two dimensional simplex, which has sufficient statistics $A = (0, 1, 2)$ and reference measure $(1, 2, 1)$. **B.** In this case $A = \{(1, 0, 0, 0), (0, 1, -1, 0)\}$ and $A_{x=4}$ is contained in the relative interior of a facet of $\mathrm{cs}(\mathcal{E})$. Here $\mathcal{E}$ is $\mathrm{Mixt}(\mathcal{E}_{\{1,2\}}, \{\delta_3\})$. **C.** In this case $A = \{(1, 0, 0, 1), (0, 1, -1, 0)\}$, such that $A_{x=4} = A_{x=1}$. All A_x are vertices of $\mathrm{cs}(\mathcal{E})$, $\mathrm{cs}(\mathcal{E})$ is a simplex.

Example 2.4.6. In Chapter 1 we showed that if $|\mathcal{X}| = 4$ ($|\mathcal{X}| = 5$) and $\mathrm{cs}(\mathcal{E})$ is a tetragon (a pentagon), then $\mathrm{Mixt}^2(\mathcal{E}) = \mathcal{P}$. In combination with Proposition 2.4.2 we get:

If $|\mathcal{X}| = 5$ ($|\mathcal{X}| = 6$) and $\mathrm{cs}(\mathcal{E})$ is a pyramid with basis given by a tetragon (a pentagon), then $\mathrm{Mixt}^2(\mathcal{E}) = \mathrm{conv}(\mathcal{E}) = \mathcal{P}$.

[1] Given an exponential family $\mathcal{E}(\mathcal{X})$, P is a kernel distribution iff there exists some $Q \in \overline{\mathcal{P}}(\mathcal{X}\setminus\mathrm{supp}(P))$ such that $P - Q \in \mathcal{N}$.

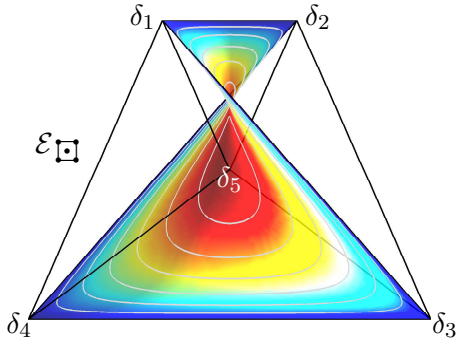

Figure 2.5: *This figure shows an exponential family on* $\{1, \ldots, 5\}$ *for which* $A_5 \in \mathrm{ri}(\mathrm{cs}(\mathcal{E}))$ *and* $\mathrm{cs}(\mathcal{E})$ *is a tetragon. The color from blue to red corresponds to values of* $p(5)$ *from* 0 *to* $\frac{1}{5}$*. The contour lines are equally spaced values of* $p(5)$*. Note that only the boundary of* $\mathcal{E}$ *intersects* $\overline{\mathcal{P}}(\{1,2,3,4\})$.

Carathéodory's theorem [24] (Steinitz' extension) states that $\mathrm{Car}(V) \leq d+1$ for any set $V \subset \mathbb{R}^d$. Hanner and Rådström [52] extended a result by Fenchel [43] and showed that if $V \subset \mathbb{R}^d$ is compact and has at most d *convexly connected*[2]components (e.g., V consists of at most d connected components), then $\mathrm{Car}(V) \leq d$. It is natural to ask for further conditions on the set V which guarantee lower Carathéodory numbers. For example, $\mathrm{Car}(S^{d-1}) = 2 \ \forall d \geq 1$. Space curves are curves in $\mathbb{R}^3$. Their Carathéodory number is at most 3. Their convex hulls are interesting objects (see [94] for a recent work on boundaries of space curves). We show the following:

Proposition 2.4.7. *Consider a continuous curve* $c\colon S^1 \to \mathbb{R}^3$*. If there exists a point* $p \in \mathrm{ri}(\mathrm{conv}(c))$ *with* $\mathrm{Car}_c(p) = 3$*, and if there exists an axis* $W := p + \mathbb{R}w$*,* $p, w \in \mathbb{R}^3$ *such that the winding function of* c *around* W *is monotone, then* $\mathrm{conv}(c)$ *has a face with Carathéodory number* 3*.*

Proof. The cone with apex p and base $2p-c$ is the set of all rays $G_p := p+\mathbb{R}_+(p-c(t))$, $t \in S^1$. G_p doesn't intersect c at any point (otherwise $\mathrm{Car}_c(p) \leq 2$), and divides $\mathbb{R}^3$ into two regions $G_p^{\pm}$ which are orthogonally convex[3]with respect to w (otherwise the winding function wouldn't be monotone). W.l.o.g. $c \subset G_p^-$ and $p + w \in G_p^+$. We have that $G_{p+\lambda w} \subset \overline{G}_p^+ \ \forall \lambda \geq 0$. This implies that $\mathrm{Car}_c(p + \lambda w) \geq 3 \ \forall \lambda \geq 0$. Clearly, there exists some $\lambda \geq 0$ for which $(p + \lambda w) \in \partial \mathrm{conv}(c)$. □

In particular, any space curve c which can be projected into the boundary of a planar convex set with regular values of degree one has Carathéodory number bounded from above by the maximal Carathéodory number of its convex hull's faces. If none of the faces contains a polygon, then $\mathrm{Car}(c) = 2$.

Corollary 2.4.8. *Let* $\mathcal{E}_{\pentagon}$ *be the exponential family from Proposition 1.2.8. Then* $\mathrm{Car}(\mathcal{E}_{\pentagon}) \leq \mathrm{Car}_{\partial\mathcal{E}_{\pentagon}}(\partial\mathcal{P}) = 2$.

[2]A set $W \subset \mathbb{R}^d$ is *convexly connected* if there is no $(d-1)$-plane U which divides W into two non-empty parts and for which $U \cap W = \emptyset$. E.g. a collection of concentric spheres is convexly connected. See [52].

[3]A subset $\mathcal{M}$ of $\mathbb{R}^d$ is orthogonally convex with respect to a direction $w \in \mathbb{R}^d$ if whenever $r, s \in \mathcal{M}$ satisfy $r - s = cw$ for some $c \in \mathbb{R}$, then $\mathcal{M}$ contains the convex hull of r and s.

Proof. Extend Proposition 2.4.7 to the case of curves in $\mathbb{R}^4$ and use that any point in $\partial\mathcal{P}$ is the mixture of two points in $\partial\mathcal{E}_{\pentagon}$. □

Example 2.4.9. Consider $\mathcal{E}_{\boxdot}$, an exponential family on $\{1,\dots,5\}$ with sufficient statistics A such that $A_{x=5}$ is contained in $\mathrm{ri}(\mathrm{cs}(\mathcal{E}))$. Then $\dim(\mathrm{Mixt}^2(\mathcal{E})) = 4$.

Proof. Consider the function $f\colon \mathcal{P} \to \mathbb{R};\ p \mapsto p_5$. The level set $\mathcal{E}^h := (f|_{\mathcal{E}})^{-1}(h)$ is the intersection of $\mathcal{E}$ with the three-dimensional affine space $\{p\colon \sum_{i=1}^5 p_i = 1, p_5 = h\}$. The moment map π maps $\mathcal{E}^h$ bijectively onto $Q^h := \pi(\mathcal{E}^h)$. □

An interesting property of minimal surfaces is that they are contained in the convex hull of their boundary. We have seen in Example 2.1.8, that $\mathcal{E}^1_{2,\mathrm{bin}}$ is not a minimal surface. On the other hand, any exponential family which contains all point measures in its closure certainly is contained in the convex hull of its boundary. The following provides sufficient conditions as well as necessary conditions:

Proposition 2.4.10. *Consider an exponential family $\mathcal{E}$ with sufficient statistics A.*

- *If* $\mathrm{cl}(\mathrm{conv}(\mathcal{E})) = \mathrm{conv}(\partial\mathcal{E})$*, then* $A_x \in \partial\,\mathrm{cs}(\mathcal{E})\ \forall x \in \mathcal{X}$.
- *If $\mathcal{E}$ contains a convex set, or if* $A_x \in \mathrm{ex}(\mathrm{cs}(\mathcal{E}))\ \forall x \in \mathcal{X}$*, then* $\mathrm{cl}(\mathrm{conv}(\mathcal{E})) = \mathrm{conv}(\partial\mathcal{E})$.
- *The condition* $A_x \in \partial\,\mathrm{cs}(\mathcal{E})\ \forall x \in \mathcal{X}$ *does not imply* $\mathrm{cl}(\mathrm{conv}(\mathcal{E})) = \mathrm{conv}(\partial\mathcal{E})$.

Proof. (i) If $A_x \in \mathrm{ri}(\mathrm{cs}(\mathcal{E}))$, then the function $\eta \mapsto p_\eta(x)$ has a unique maximum $\eta = F \cdot \delta_x = A(x)$. To see this, consider the derivative $\partial_\eta p_\eta(x) = (A(x) - \eta)p_\eta \partial_\eta\theta$, and note that $\eta = A(x)$ is the only critical point which comes into question. For $\eta = A(x)$, $p_\eta \notin \mathrm{conv}(\partial\mathcal{E})$.

(ii) If $\mathcal{E}$ contains a convex set, then $\mathcal{E}$ is a ruled manifold, and every $p \in \mathcal{E}$ is contained in the straight line connecting two points in $\partial\mathcal{E}$. If $A_x \in \mathrm{ex}(\mathrm{cs}(\mathcal{E}))\ \forall x$ and $A_x \neq A_y\ \forall x \neq y$ then $\{\delta_x\}_{x\in\mathcal{X}} \subset \overline{\mathcal{E}}$. If $A_x \in \mathrm{ex}(\mathrm{cs}(\mathcal{E}))\ \forall x$ and $A_x = A_y$ for some $x \neq y$, then there is a partition $\varrho = \{\mathcal{X}_y\}_y$ of $\mathcal{X}$, where $\mathcal{X}_y := \{x \in \mathcal{X}\colon A_x = A_y\}$. We have that $\mathcal{E}$ is a subfamily of $\mathcal{S}_\varrho$ and $\overline{\mathcal{E}} \supset \mathrm{ex}(\mathcal{S}_\varrho)$. *(iii)* See Example 2.4.11 and Figure 2.6. □

Example 2.4.11. Consider the two-dimensional exponential family $\mathcal{E}$ on $\mathcal{X} = \{1,\dots,5\}$ with sufficient statistics $A = \begin{pmatrix} 1 & 1 & \frac{1}{2} & 0 & 1 \\ 0 & 1 & \frac{1}{2} & 1 & \frac{1}{2} \end{pmatrix}$ and uniform reference measure. The configuration of A_x is as $\bigtriangleup$, such that $A_x \in \partial\,\mathrm{cs}(\mathcal{E})\ \forall x \in \mathcal{X}$. The facial sets are $\mathcal{F}(\mathcal{E}) = \{\{12345\}, \{125\}, \{134\}, \{24\}, \{1\}, \{2\}, \{4\}\}$. Figure 2.6 shows a linear image of $\mathcal{E}$ and of the convex hull of its boundary. The later doesn't contain all points of $\mathcal{E}$. This confirms item three of Proposition 2.4.10.

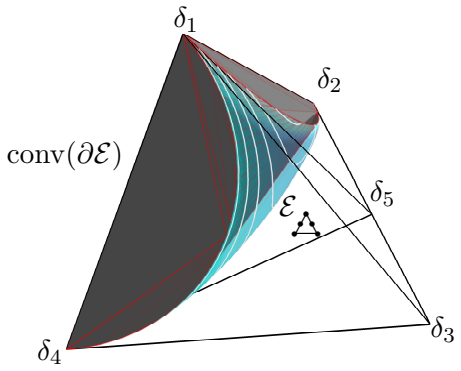

Figure 2.6: *This figure illustrates Example 2.4.11. Depicted is a linear projection of the probability simplex on $\mathcal{X} = \{1, \ldots, 5\}$ onto a three-dimensional simplex. The light blue surface is the projection of a two-dimensional exponential family on $\mathcal{X}$ for which all columns of its sufficient statistics are in the boundary of the convex support, but only three of them are also vertices. The white curves are level surfaces of $p(5)$. The gray volume is the projection of the convex hull of the boundary of $\mathcal{E}$. The red lines show a triangulation of the boundary of this volume. This figure shows $\mathrm{conv}(\partial\mathcal{E}) \not\supset \mathcal{E}$.*

Part II.
Restricted Boltzmann Machines and Deep Belief Networks

3 Universal Approximation Results for RBMs and DBNs

Restricted Boltzmann Machines

A *Boltzmann Machine* is an undirected stochastic binary network, formally similar to the Ising model in statistical physics. A Boltzmann Machine includes pair interactions between any two nodes, and a bias term for each node (corresponding to an external field). See [2, 59, 9, 68] for an overview. A *Glauber dynamics* can be defined on the states of the nodes. The state of each node is updated asynchronously and takes value 0 or 1 with a probability that depends on the state of its neighbors, the strength of the connection between them, and a bias term. A Restricted Boltzmann Machine (RBM) is a special type of Boltzmann Machine, where the graph describing the interactions is bipartite: Only connections between a *visible* and *hidden* part of the units appear (see Figure 3.1). An arbitrary weight can be assigned to each edge and to each unit. An RBM with n visible and m hidden units generates stationary probability distributions on the states of the visible units which have the following form:

$$p_{W,C,B}(v) = \frac{1}{Z_{W,C,B}} \sum_{h \in \{0,1\}^m} \exp\left(h^\top W v + C^\top h + B^\top v\right) \quad \forall v \in \{0,1\}^n \,, \tag{3.1}$$

where $h \in \{0,1\}^m$ denotes the state vector of the hidden units, $W \in \mathbb{R}^{m \times n}, C \in \mathbb{R}^m$ and $B \in \mathbb{R}^n$ constitute the model parameters, and $Z_{W,C,B}$ is a normalization constant (the *partition function*). A Restricted Boltzmann Machine model with n visible and m hidden units, denoted $\mathrm{RBM}_{n,m}$, is the set of all probability distributions on $\{0,1\}^n$ which can be approximated arbitrarily well by a probability distribution of the form given in eq. (3.1). In particular, the model $\mathrm{RBM}_{n,m}$ is the closure of a hierarchical model marginalized over m variables. See [107, 45] for the origins of RBMs and [72, 57, 87, 88] for additional details.

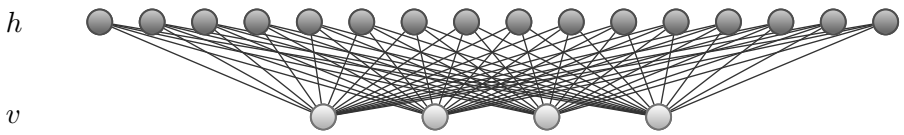

Figure 3.1: *Graphical representation of an RBM with* 4 *visible (light gray) and* 16 *hidden units (dark gray). This model can approximate any distribution on* $\{0,1\}^4$ *arbitrarily well as the stationary distribution on the states of its visible units through appropriate choice of parameters (this follows from a result from* [72]). In Section 3 we show that the number of hidden units can be halved.

Deep Belief Networks

A Deep Belief Network (DBN) is a special kind of generative graphical model with binary variables, originally introduced in 2006 by G. E. Hinton, S. Osindero and Y. Teh [58]. The graphical representation consists of a sequence of layers of units, where all pairs of units from neighboring layers are connected, but units in the same layer are not connected. The two layers at the top of the network have undirected connections between them, while all other connections are directed towards the bottom layer, which is the visible layer, see Figure 3.2. An arbitrary weight can be assigned to every edge. Beside the connection weights, every node contains an individual bias weight. Formally a DBN with only two layers is just an RBM, but the general idea is that DBNs have several hidden layers. A model with the same interaction graph as a DBN but with undirected connections is called a *Deep Boltzmann Machine* (DBM), see [99].

A DBN is specified by the number of hidden layers $l \in \mathbb{N}$ (indexed by $k \in [l]$), the number of units $n_k \in \mathbb{N}$ in the layer k, called the *width* of the layer, for each $k \in [l]$, and the width $n_0 \in \mathbb{N}$ of the visible layer. The model contains a total of $N = \sum_{k=0}^{l} n_k$ binary units. To any pair of units j and i belonging to a pair of subsequent layers $(k-1)$ and k the associated connection weight is denoted $W_{j,i}^k \in \mathbb{R}$. To any unit j in any layer k the associated bias weight is denoted $b_j^k \in \mathbb{R}$. This makes a total of $d = (\sum_{k=1}^{l} n_{k-1}n_k) + (\sum_{k=0}^{l} n_k)$ parameters. The state of the unit j in the layer k is denoted by $h_j^k \in \{0,1\}$, and the states of all units in layer k by $(h_j^k)_j =: h^k \in \{0,1\}^{n_k}$. The joint probability distributions on the states of all N units of the model are parametrized by the following map $Q : \mathbb{R}^d \to \mathcal{P}_N \subset \mathbb{R}^{2^N}$, which maps the connection weights $\{(W_{j,i}^k)_{j,i} =: W^k \in \mathbb{R}^{n_{k-1}\times n_k}\}_{k=1}^{l}$ and the bias weights $\{(b_j^k)_j =: b^k \in \mathbb{R}^{n_k}\}_{k=0}^{l}$ with a distribution P defined as follows [110]:

$$P(h^0, h^1, \ldots, h^l) = P(h^{l-1}, h^l) \prod_{k=0}^{l-2} P(h^k | h^{k+1}) \,, \tag{3.2}$$

$$P(h^k | h^{k+1}) = \prod_{j=1}^{n_k} P(h_j^k | h^{k+1}) \,, \tag{3.3}$$

$$P(h_j^k | h^{k+1}) \propto \exp\left(h_j^k b_j^k + h_j^k \sum_{i=1}^{n_{k+1}} W_{j,i}^{k+1} h_i^{k+1} \right) \,. \tag{3.4}$$

The closure of $Q(\mathbb{R}^d) \subseteq \mathcal{P}_N \subset \mathbb{R}^{2^N}$ contains all distributions which can be approximated arbitrarily well by a distribution of the form given above. We denote this set by $\mathcal{D}(n_0^l)$.

The set of probability distributions which can be approximated arbitrarily well by the DBN model, denoted $\mathrm{DBN}_{n_0,n_1,\ldots,n_l}$ or $\mathrm{DBN}(n_0^l)$ for short, is the image of $\mathcal{D}(n_0^l)$ by the linear marginal map:

$$\begin{aligned} \mathcal{D}(n_0^l) &\to \mathrm{DBN}(n_0^l) \subseteq \mathcal{P}_n \,, \\ P &\mapsto p = M \cdot P \,, \end{aligned}$$

where $M \in \mathbb{R}^{2^n \times 2^N}$ has rows $M_v = \mathbb{1}_{\{(v',h'):v'=v\}}$ $\forall v \in \{0,1\}^n$. This is just $p(v) = \sum_{h\in\{0,1\}^{N-n}} P(v,h)$.

The model $\mathrm{DBN}(n_0^l)$ only depends on the ordered tuple $n_0, \ldots, n_l$. For notational convenience we sometimes use subscripts from an interval $\{r, r+1, \ldots, r+l\} \subset \mathbb{Z}$. A DBN is

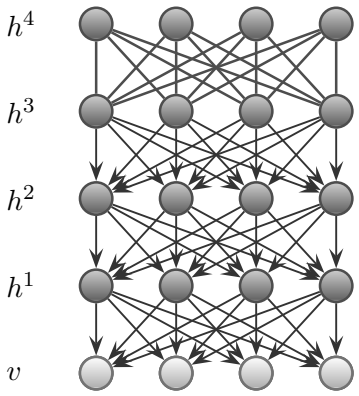

Figure 3.2: *This figure shows the graph of interactions of a narrow DBN with* 4 *visible units (at the bottom layer). A DBN with this architecture can approximate any distribution on* $\{0,1\}^4$ *arbitrarily well (this follows from a result from* [73]). In Section 3.2 we improve previous bounds on the minimal number of layers of a DBN universal approximator (we show that roughly half the number of hidden layers used in previous results suffice).

narrow if all hidden layers have width of order n_0, the width of the visible layer.

Universal Approximation

An RBM, respectively a DBN which can approximate any distribution on the states of its visible units arbitrarily well is called a *universal approximator*. We say that an RBM, DBN, or some other statistical model can approximate some probability distribution $p \in \overline{\mathcal{P}_n}$ arbitrarily well iff there exists a sequence of visible probability distributions in the model which converges to p. The main question in this chapter is: *What is the minimal number of hidden units (hidden layers) of an RBM (a narrow DBN) universal approximator?* This chapter builds on previous work by N. Le Roux and Y. Bengio [72, 73]. We improve their results on the minimal size of RBM and DBN universal approximators and resolve thereby a conjecture that they formulated. The results of this chapter are used in Chapter 4, where we analyze the expressive power of RBMs and DBNs which are not necessarily universal approximators.

A DBN universal approximator on $\{0,1\}^n$ necessarily has a number of parameters larger or equal to $2^n - 1$, the dimension of $\mathcal{P}_n$. We give a formal proof of this intuitive statement in Appendix 3.A. A lower bound on the number of hidden layers of a DBN universal approximator with layers of width n is $\frac{2^n-1-n}{n(n+1)}$. This implies the answer *yes* to the following question, which was raised in [110]: *Given that a network with* $2^n/n^2$ *layers has about* 2^n *parameters, can it be shown that a deep and narrow (with width* $n + c$*) network of* $\ll 2^n/n^2$ *layers cannot approximate every distribution?*

Using Corollary 1.3.3 from Section 1, a further bound can be given as:

Proposition 3.0.12. *Any RBM, DBN, or DBM universal approximator on* $\{0,1\}^n$ *has at least* $(n-1)$ *units in the first hidden layer (next to the visible layer).*

Proof. The visible distributions are mixtures of 2^{n_1} product distributions, where n_1 is the number of units in the first hidden layer. By Corollary 1.3.3, the mixture can't represent distributions supported by $Z_{\pm,n}$ unless $2^{n_1} \geq 2^{n-1}$. □

Since DBNs and RBMs have restricted architectures, the lower bounds derived above are not necessarily attained by universal approximators of these kinds.

The following theorem shows that RBMs are universal approximators provided they have enough hidden units (see also [45]):

Theorem 3.0.13. (N. Le Roux and Y. Bengio [72, Theorem 2]). *Any distribution on $\{0,1\}^n$ with support of cardinality s can be approximated arbitrarily well (with respect to the Kullback-Leibler divergence) by an RBM with $(s+1)$ hidden units.*

I. Sutskever and G. E. Hinton [110] showed the existence of narrow DBN universal approximators. More precisely, they showed that a DBN with $\sim 3 \cdot 2^n$ hidden layers of width $(n+1)$ can approximate any distribution on $\{0,1\}^n$ arbitrarily well. In [73] it is shown that hidden layers of width n suffice, and furthermore:

Theorem 3.0.14. (N. Le Roux and Y. Bengio [73, Theorem 4]). *If $n = 2^t$, a DBN composed of $\frac{2^n}{n} + 1$ layers of size n is a universal approximator of distributions on $\{0,1\}^n$.*

The optimality of [73, Theorem 4] remains an open problem in that paper. However, the proof method suggests the sufficiency of less hidden layers (of order $\frac{2^n}{2n}$), which was conjectured in that paper. Our Theorem 3.2.1 gives a positive solution to that conjecture. The proofs contained in [73] crucially depend on Theorem 3.0.13. In this chapter we sharpen that ingredient, see Theorem 3.1.1, which allows us to even better exploit their method, see Lemma 3.2.3, and thereby prove Theorem 3.2.1.

3.1 Restricted Boltzmann Machines

The following Theorem 3.1.1 improves Theorem 3.0.13.

Theorem 3.1.1. (RBM universal approximators). *Any distribution p on binary vectors of length n can be approximated arbitrarily well by an RBM with $(k-1)$ hidden units, where k is the minimal number of pairs of binary vectors such that the two vectors in each pair have Hamming distance one and such that the support set of p is contained in the union of these pairs.*

We shall present a stronger version of this result in Chapter 4 (Theorem 4.2.1).

Any subset of $\{0,1\}^n$ can be covered by 2^{n-1} pairs of vectors of Hamming distance one, because the graph of the n-dimensional hypercube (the graph of the cube has a perfect matching). The minimal number of pairs which is sufficient to cover the support of some $p \in \overline{\mathcal{P}_n}$ can be as small as $|\operatorname{supp}(p)|/2$. For example, if the support of p is of the form $\{(x_1,\ldots,x_n) : x_{i_j} \in \{0,1\}, 1 \leq j \leq b\} \subseteq \{0,1\}^n$ for any $1 \leq b \leq n$ and fixed $x_i \in \{0,1\}$ for all $i \neq i_1,\ldots,i_b$. The union of the following 2^{b-1} pairs covers that set:

$$\left\{\{(x_1,\ldots,\underset{i_1}{0},\ldots,x_n),(x_1,\ldots,\underset{i_1}{1},\ldots,x_n)\} : x_{i_j} \in \{0,1\}, 2 \leq j \leq b\right\}.$$

Therefore, we have the following:

Corollary 3.1.2.

- *Any distribution on* $\{0,1\}^n$ *can be approximated arbitrarily well by an RBM with* $\frac{2^n}{2} - 1$ *hidden units.*
- *An RBM with* n *hidden units can approximate* $p \in \overline{\mathcal{P}_n}$ *arbitrarily well, whenever* $\mathrm{supp}(p)$ *is contained in the set of vertices of some* $(\log(2(n+1)))$*-dimensional face of the* n*-dimensional unit cube, e.g.,* $\mathrm{supp}(p) = \{(x_1, \ldots, x_b, 0, \ldots, 0) \in \{0,1\}^n : x_i \in \{0,1\}, 1 \leq i \leq b\}$ *for any* $b \leq \log(2(n+1))$.

Our proof is in the spirit of the proof of [72, Theorem 2]. The idea of that proof is to show that, given an RBM with some marginal visible distribution, appending one additional hidden unit allows to increment the probability mass of one visible state vector, while uniformly reducing the probability mass of all other visible state vectors. We show that appending an additional hidden unit in fact allows to increase the probability mass of a pair of visible vectors, in independent ratio, given that this pair differs in one entry. At the same time, the probability of all other visible states is reduced uniformly. Furthermore, we use the bias weights in the visible layer to improve the result.

Proof of Theorem 3.1.1. (i) Let p be the distribution on the states of visible and hidden units of an RBM that is represented for a choice of the parameters W, B and C. Its marginal distribution on v can be written as

$$p(v) = \frac{\sum_h z(v,h)}{\sum_{v',h'} z(v',h')} \,, \tag{3.5}$$

where $z(v,h) = \exp(hWv + Bv + Ch)$. Denote by $p_{w,c}$ the distribution that arises when an additional hidden unit is added to the RBM connected with weights $w = (w_1, \ldots, w_n)$ to the visible units, and with bias weight c. Its marginal distribution is

$$p_{w,c}(v) = \frac{(1 + \exp(w \cdot v + c)) \sum_h z(v,h)}{\sum_{v',h'} (1 + \exp(w \cdot v' + c)) z(v',h')} \,. \tag{3.6}$$

(ii) Given any vector $v \in \{0,1\}^n$ we write $v^{j,0}$ for the vector defined through $v_i^{j,0} = v_i, \forall i \neq j$, and $v_j^{j,0} = 0$. Similarly we write $v^{j,1}$ for the vector with $v_i^{j,1} = v_i, \forall i \neq j$, and $v_j^{j,1} = 1$. We also write $\mathbb{1} := (1, \ldots, 1)$, and $\mathbf{e}_j := (0, \ldots, 0, 1, 0, \ldots, 0)$, where the 1 is in the j-th entry.

(iii) Consider an arbitrary $j \in \{1, \ldots, n\}$ and an arbitrary visible vector u. Consider also $s := \mathbb{1} \cdot u^{j,0}$, i.e., the number of ones in vector $u^{j,0}$. Define

$$\begin{aligned} \hat{w} &:= a(u^{j,0} - \frac{1}{2}\mathbb{1}^{j,0}) \,, \\ \bar{w} &:= \hat{w} + (\lambda_2 - \lambda_1)\mathbf{e}_j \,, \\ \bar{c} &:= -\hat{w} \cdot u^{j,0} + \lambda_1 = -\hat{w} \cdot u^{j,1} + \lambda_1 \,. \end{aligned}$$

For the weights $\bar{w}$ and $\bar{c}$ we have:

$$\bar{w} \cdot v = \frac{1}{2}a(s - |\{i : u_i^{j,0} \neq v_i^{j,0}\}|) + (\lambda_2 - \lambda_1)v_j \,, \tag{3.7}$$

$$\bar{c} = -\frac{1}{2}as + \lambda_1 \,, \tag{3.8}$$

and in the limit $a \to \infty$ we get:

$$\begin{aligned} \lim_{a\to\infty} 1 + \exp(\bar{w} \cdot v + \bar{c}) &= 1, \quad \forall v \neq u^{j,1}, u^{j,0} , \\ \lim_{a\to\infty} 1 + \exp(\bar{w} \cdot u^{j,0} + \bar{c}) &= 1 + e^{\lambda_1} , \\ \lim_{a\to\infty} 1 + \exp(\bar{w} \cdot u^{j,1} + \bar{c}) &= 1 + e^{\lambda_2} . \end{aligned} \tag{3.9}$$

Now we take a look at the denominator in the right hand side of eq. (3.6). For the parameters $\bar{w}$ and $\bar{c}$ defined above this evaluates to:

$$\lim_{a\to\infty} \sum_{v',h'} (1 + \exp(\bar{w} \cdot v' + \bar{c}))z(v', h') = \sum_{v',h'} z(v', h') + e^{\lambda_1} \sum_{h'} z(u^{j,0}, h') + e^{\lambda_2} \sum_{h'} z(u^{j,1}, h') . \tag{3.10}$$

Inserting the terms of eqs. (3.9) and eq. (3.10) into eq. (3.6), and multiplying nominator and denominator by $\sum_{v',h'} z(v', h')$ yields (cf. eq. (3.5)):

$$\begin{aligned} \lim_{a\to\infty} p_{\bar{w},\bar{c}}(v) &= \frac{p(v)}{1 + e^{\lambda_1} p(u^{j,0}) + e^{\lambda_2} p(u^{j,1})} , \quad \forall v \neq u^{j,1}, u^{j,0} , \\ \lim_{a\to\infty} p_{\bar{w},\bar{c}}(u^{j,0}) &= \frac{(1 + e^{\lambda_1})p(u^{j,0})}{1 + e^{\lambda_1} p(u^{j,0}) + e^{\lambda_2} p(u^{j,1})} , \\ \lim_{a\to\infty} p_{\bar{w},\bar{c}}(u^{j,1}) &= \frac{(1 + e^{\lambda_2})p(u^{j,1})}{1 + e^{\lambda_1} p(u^{j,0}) + e^{\lambda_2} p(u^{j,1})} . \end{aligned} \tag{3.11}$$

This shows that the probability of $u^{j,0}$ and of $u^{j,1}$ can be increased independently by a multiplicative factor, while all other probabilities are reduced uniformly.

(iv) Now we explain how to start an induction from which the claim follows. Consider an RBM with no hidden units. Through a choice of the bias weights in every visible unit, $\mathrm{RBM}_{n,0}$ produces any arbitrary product distribution $p^0(v) \propto \exp(B \cdot v) \propto \exp(B \cdot v + K)$ as visible distribution, where B is the vector of bias weights and K is a constant that we introduce for illustrative reasons, and is not a parameter of the $\mathrm{RBM}_{n,0}$ since it cancels out with the normalization of p^0. In particular, $\mathrm{RBM}_{n,0}$ can approximate arbitrarily well any distribution with support given by a pair of vectors that differ in only one entry. To see this, consider any pair of vectors $u^{j,0}$ and $u^{j,1}$ that differ in the entry j. Then the choice $B = a(u^{j,0} - \frac{1}{2}\mathbb{1}^{j,0}) + (\lambda_2 - \lambda_1)\mathbf{e}_j$ and $K = -a(u^{j,0} - \frac{1}{2}\mathbb{1}^{j,0})u^{j,0} + \lambda_1$ yields in the limit $a \to \infty$ (similarly to eqs. (3.9)) that $\lim_{a\to\infty} p^0(v) = 0$ whenever $v \neq u^{j,1}$ and $v \neq u^{j,0}$, while $\lim_{a\to\infty} p^0(u^{j,1})/p^0(u^{j,0}) = \exp(\lambda_2 - \lambda_1)$ can be chosen arbitrarily by modifying λ_1 and λ_2. Hence p^0 can be made arbitrarily similar to any distribution with support $\{u^{j,1}, u^{j,0}\}$. Notice that p^0 remains always strictly positive for $a < \infty$.

By the arguments described above in eqs. (3.11), every additional hidden unit allows to increase the probability of any pair of vectors which differ in one entry. Obviously, it is possible to do the same for a single vector instead of a pair. Hence $\mathrm{RBM}_{n,(i-1)}$ is an approximator of distributions with support contained in any union of i pairs of vectors which differ in exactly one entry. □

3.2 Deep Belief Networks

In this section we make a sensible modification of the construction used in the proof of [73, Theorem 4] and prove Theorem 3.2.1, the main result of this chapter:

Theorem 3.2.1. (DBN universal approximators). *Let $b \in \mathbb{N}$ and $n = \frac{2^b}{2} + b$. A DBN containing $\frac{2^n}{2(n-b)}$ hidden layers of width n is a universal approximator of distributions on $\{0,1\}^n$.*

We first develop some components of the proof. An important idea of [110] is that of *sharing*, by means of which in a part of a DBN the probability of a vector is increased while the probability of another vector is decreased and the probability of all other vectors remains nearly constant. This idea is refined in [73, Theorem 2]:

Consider two layers of units indexed by $i \in [n_1]$ and $k \in [n_2]$. Denote $\{w_{ik}\}_{i,k}$ the connection weights and $\{c_i\}$ the bias weights in the first layer. Denote $v \in \{0,1\}^{n_1}$ and $h \in \{0,1\}^{n_2}$ the state vectors for each layer. Let $a, b \in \{0,1\}^{n_2}$ be vectors satisfying $d_H(a,b) = 1$ and let $j \in [n_2]$ be the entry where they differ.

Theorem 3.2.2. (N. Le Roux and Y. Bengio [73, Theorem 2]). *Given any $l \in [n_1]$ there exist weights $\{w_{l,k}\}_{k\in[n_2]}$ and c_l such that the following equations are satisfied with arbitrary accuracy: $P(v_l = h_l|h) = 1\ \forall h \notin \{a,b\}$, while $P(v_l = 1|h = a) = p_a$ and $P(v_l = 1|h = b) = p_b$ with arbitrary $p_a, p_b \in [0,1]$.*

By this theorem, a sharing step can be accomplished in only one layer, where probability mass is transferred from a chosen vector to another vector that differs in one entry. The sharing step requires only adaptation of the connection weights and bias weight of one single unit. This way, the overlay of a number of sharing steps in each layer is possible. The requirements for the realizability of simultaneous sharing steps as described in Theorem 3.2.2 can be summarized in properties of sequences of binary vectors. These properties are described in [73, Theorem 3], or in the items 2–3 of the following lemma:

Lemma 3.2.3. *Let $b \in \mathbb{N}$, $n = \frac{2^b}{2} + b$, and $a := 2(n-b) = 2^b$. There exist 2^b sequences of binary vectors S_i, $0 \le i \le a-1$ composed of vectors $S_{i,k}, 1 \le k \le \frac{2^n}{a}$ satisfying the following:*

1. *$\{S_0, \ldots, S_{a-1}\}$ is a partition of $\{0,1\}^n$.*
2. *$d_H(S_{i,k}, S_{i,k+1}) = 1\ \forall i \in \{0,\ldots,a-1\}, \forall k \in \{1,\ldots,\frac{2^n}{a}-1\}$.*
3. *For all $i, j \in \{0,\ldots,a-1\}$, $i \neq j$ and every $k \in \{1,\ldots,\frac{2^n}{a}-1\}$ the bit switched between $S_{i,k}$ and $S_{i,k+1}$ and the bit switched between $S_{j,k}$ and $S_{j,k+1}$ are different, unless $d_H(S_{i,k}, S_{j,k}) = 1$.*
4. *Furthermore, the choice $\{S_{0,1}, \ldots, S_{a-1,1}\} = \{(x, 0, \ldots, 0) : x \in \{0,1\}^b\}$, (the vertices of a b-dimensional face of the n-cube), is possible.*

Proof of Lemma 3.2.3. An m-bit Gray code is a matrix of size $2^m \times m$, where every element from $\{0,1\}^m$ appears exactly once as a row of the matrix, and any two consecutive rows have Hamming distance one to each other. A Gray code can be understood as an ordered binary code describing a Hamiltonian path on the graph of the m-cube. Such paths exist for any m, (the graph of the m-cube is Hamiltonian). Let G^0_{n-b} be any $(n-b)$-bit Gray code. Obviously, any

permutation of columns of a Gray code gives a Gray code. Let G^i_{n-b} be the cyclic permutation of the columns of G^0_{n-b}, i positions to the left. Now, we define the following matrix:

$$S_i = \begin{pmatrix} S_{i,1} \\ \vdots \\ S_{i,k} \\ \vdots \\ S_{i,2^{n-b}} \end{pmatrix} := \begin{pmatrix} \mathrm{bin}_b(i) & \\ \vdots & G^{i \bmod (n-b)}_{n-b} \\ \mathrm{bin}_b(i) & \end{pmatrix} .$$

The first b entries of the row vector $S_{i,k}$ contain the b-bit representation of i. The remaining $(n-b)$ entries of $S_{i,k}$ contain the k-th row of the Gray code $G^{i \bmod (n-b)}_{n-b}$, which is G^0_{n-b} with columns cyclically shifted i positions to the left. The cyclic shift forces two sequences of vectors S_i and S_j, $i \neq j$ to change the same bit in the same row only if $G^{i \bmod (n-b)}_{n-b} = G^{j \bmod (n-b)}_{n-b}$ (in this case both sequences change the same bit in every row), i.e., only if $i = j \mod (n-b)$, which means that i and j differ in a multiple of $(n-b) = 2^b/2$. This implies that $\mathrm{bin}_b(i)$ and $\mathrm{bin}_b(j)$ differ in exactly the first entry. For the last item, note that G^0_{n-b} can be chosen such that the first row is $(0, \ldots, 0)$. Thus we have verified all claims. □

Remark 3.2.4. Two consecutive rows $S_{i,k}$ and $S_{i,k+1}$ in a sequence S_i differ in an entry that can be located in almost any position $\{1, \ldots, n\}$. In contrast, in the sequences from [73, Theorem 3] that entry can be located only in a subset of $\{1, \ldots, n\}$ of cardinality $n/2$. In Lemma 3.2.3, each of $(n-b)$ entries of any row is flipped by exactly two sequences. The choice of the relations between number of sequences, number of visible units, and number of layers is not accidental and somewhat intricate. It must take into account all the components that will be needed in the proof of Theorem 3.2.1. The attempt to produce $2n$ instead of $2(n-b)$ sequences with the properties 1 and 2 (and flips in all entries) would correspond to the following: Set $\begin{pmatrix} S_0 \\ \vdots \\ S_{2n-1} \end{pmatrix} = G_n$, i.e., the sequences to be overlayed are portions of the same Gray code. In this case it is difficult to satisfy property 3, i.e., that if S_i and S_j flip the same bit in the same row, then $d_H(S_{i,k}, S_{j,k}) = 1$. This property however is essential for using Theorem 3.2.2. Most common Gray codes flip some entries more often than other entries and can be discarded. Other sequences referred to as *totally balanced Gray codes* flip all entries equally often and exist whenever n is a power of 2, but still a strong cyclicality condition would be required for our purposes. On account of this we say that the sequences given in our Lemma 3.2.3 allow optimal use of [73, Theorem 2].

The following Lemma 3.2.5 is a transcription of [73, Lemma 1] with replacements of indices according to our construction. Denote by h^i a state vector of the units in the hidden layer i, and denote by h^0 a visible state. The joint distribution on the states of all units, for $\frac{2^n}{a} + 1$ layers, is of the form $P(h^0, h^1, \ldots, h^{\frac{2^n}{a}}) = P(h^{\frac{2^n}{a}-1}, h^{\frac{2^n}{a}}) \prod_{k=1}^{\frac{2^n}{a}-1} P(h^{\frac{2^n}{a}-(k+1)} | h^{\frac{2^n}{a}-k})$.

Lemma 3.2.5. *Let p^* be an arbitrary distribution on $\{0,1\}^n$. Consider a DBN with $\frac{2^n}{a} + 1$ layers and the following properties:*

1. *For all $i \in \{0, \ldots, a-1\}$ the top RBM between $h^{\frac{2^n}{a}}$ and $h^{\frac{2^n}{a}-1}$ assigns probability $\sum_k p^*(S_{i,k})$ to $S_{i,1}$,*

2. *For all* $i \in \{0, \ldots, a-1\}$ *and for all* $k \in \{1, \ldots, \frac{2^n}{a} - 1\}$

$$P(h^{\frac{2^n}{a}-(k+1)} = S_{i,k+1} | h^{\frac{2^n}{a}-k} = S_{i,k}) = \frac{\sum_{t=k+1}^{\frac{2^n}{a}} p^*(S_{i,t})}{\sum_{t=k}^{\frac{2^n}{a}} p^*(S_{i,t})} ,$$

$$P(h^{\frac{2^n}{a}-(k+1)} = S_{i,k} | h^{\frac{2^n}{a}-k} = S_{i,k}) = \frac{p^*(S_{i,k})}{\sum_{t=k}^{\frac{2^n}{a}} p^*(S_{i,t})} .$$

3. *For all* $k \in \{1, \ldots, \frac{2^n}{a} - 1\}$

$$P(h^{\frac{2^n}{a}-(k+1)} = u | h^{\frac{2^n}{a}-k} = u) = 1, \quad \forall u \notin \cup_i \{S_{i,k}\} .$$

Such a DBN has p^* *as its marginal visible distribution.*

Now we are ready to prove Theorem 3.2.1:

Proof of Theorem 3.2.1. The proof follows the strategy of the proof of [73, Theorem 4] given in that paper. We show the existence of a DBN with the properties of the DBN described in Lemma 3.2.5. In view of Corollary 3.1.2 it is possible to achieve that the top RBM assigns arbitrary probability to a collection of vectors $S_{i,1}, i \in \{0, \ldots, a-1\}$ whenever they are contained in the set of vertices of a $\log(2(n+1))$-dimensional cube. This requirement is met for the vectors $S_{i,1}, i \in \{0, \ldots, a-1\}$ of Lemma 3.2.3, since we can choose $\{S_{i,1}\}_i = \{(x, 0, \ldots, 0) \in \{0,1\}^n : x \in \{0,1\}^b\}$, which is a b-dimensional cube, and $b < \log 2n$. At each subsequent layer, the first b bits of h^{k+1} are copied to the first b bits of h^k with probability arbitrarily close to one. The $(n-b)$ remaining bits are potentially changed to move from one vector in a Gray code sequence to the next with the correct probability as defined in Lemma 3.2.5. This changes are possible because Theorem 3.0.13 can be applied for the sequences provided in Lemma 3.2.3. The crucial difference to the proof of Theorem 3.0.14 is that by our definition of the $\{S_i\}$, at each layer $(n-b)$ bit flips occur (with correct probabilities), instead of $\frac{n}{2}$. □

3.A Lower Bounds on the Number of Parameters

In this appendix we formally confirm the heuristic that a DBN can only approximate any visible distribution on $\{0,1\}^n$ arbitrarily well when the number of parameters of that DBN is not less than $2^n - 1$.

Consider a DBN with l hidden layers, where the hidden layer $k = 1, \ldots, l$ contains n_k units. Let N be the total number of units of the DBN, and d the number of parameters: The connection weights $W_{j,i}^{k+1}$ between the unit j in layer k and the unit i in layer $k+1$, for all j and i, and $k = 0, \ldots, l-1$, as well as the biases b_j^k for all j and $k = 0, \ldots, l$.

The set of joint distributions on the states of all units of the DBN which arise through variation of the parameters is a manifold $\mathcal{E} = Q(\mathbb{R}^d) \subset \mathcal{P}_N \subset \mathbb{R}^{2^N}$ of dimension not more than d, parametrized by the function $Q : \mathbb{R}^d \to \mathcal{P}_N \subset \mathbb{R}^{2^N}$, which takes the parameters $\{W_{j,i}^k\}, \{b_j^k\}$ into a distribution P defined as in eqs. (3.2)–(3.4). Q is continuous everywhere and converges to some distribution for any sequence of parameters escaping to any direction. Hence $Q(\overline{\mathbb{R}^d})$, i.e.,

the set of all joint distributions (also for parameters taking infinite values, including not strictly positive distributions), is contained in a compact set $\overline{\mathcal{E}}$ which is contained in a bounded manifold of dimension $\dim(\mathcal{E})$.

Now, restricting observations to the visible units corresponds to marginalizing out the variables $h^k, k = 1, \ldots, l$, which is applying the following linear map:

$$\begin{aligned} \mathrm{M}: \quad \mathbb{R}^{2^N} \supset \overline{\mathcal{E}} \quad &\to \quad \overline{\mathcal{P}_n} \subseteq \mathbb{R}^{2^n} \ ; \\ p \quad &\mapsto \quad p_V = M \cdot p \ , \end{aligned}$$

for a matrix $M \in \mathbb{R}^{2^n \times 2^N}$ with rows $M_v = \mathbb{1}_{\{(v',h'):v'=v\}}$, for $v \in \{0,1\}^n$, such that

$$p_V(v) = \sum_{h \in \{0,1\}^{N-n}} p(v,h) \ .$$

Since this is a linear map, its differential (the Jacobian of the natural extension of M to $\mathbb{R}^{2^N}$ restricted to the tangential space of $\mathcal{E}$) is given by the same matrix: $d_p \mathrm{M} = M : T_p\mathcal{E} \to T_{p_V}\mathcal{P}_n$. The rank of this map is not more than $\dim T_p\mathcal{E} = \dim \mathcal{E}$. The elements $p \in \mathcal{E}$ for which the differential $d_p \mathrm{M}$ is not a surjective map are called *critical points*, and for these p the value $\mathrm{M}(p)$ is called a *critical value*. If $\dim \mathcal{E} < \dim \mathcal{P}_n = 2^n - 1$, then clearly all points in the image $\mathrm{M}(\mathcal{E})$ are critical values. Sard's theorem [101], states that the set of critical values is a null set. This means that if $\dim \mathcal{E} < \dim \mathcal{P}_n$, then $\mathrm{M}(\mathcal{E})$ is a null set of $\mathcal{P}_n$. $\mathrm{M}(\overline{\mathcal{E}})$ is also a null set, since M can be extended to a domain which is a manifold containing $\overline{\mathcal{E}}$, and the image of which is a null set of $\operatorname{aff} \mathcal{P}_n$, i.e., a null set of $\mathcal{P}_n$. Note that a set $\mathcal{G}$ approximates any element of $\mathcal{P}_n$ arbitrarily well exactly when it is dense in $\mathcal{P}_n$, i.e., $\overline{\mathcal{G}} = \overline{\mathcal{P}_n}$. Since the map M is continuous and $\overline{\mathcal{E}}$ is compact, we have that $\mathrm{M}(\overline{\mathcal{E}})$ is a compact subset of $\overline{\mathcal{P}_n}$. In particular, $\overline{\mathrm{M}(\overline{\mathcal{E}})} = \mathrm{M}(\overline{\mathcal{E}})$. By the above-arguments this is a null set whenever $\dim \mathcal{E} < \dim \mathcal{P}_n$, in which case it obviously differs from $\overline{\mathcal{P}_n}$. Hence if a DBN approximates any visible distribution on $\{0,1\}^n$ arbitrarily well, then the number of its parameters is at least equal to $2^n - 1$, the dimension of the set of all distributions on $\{0,1\}^n$. The arguments given above allow a straightforward generalization to the case of RBMs, DBMs, Boltzmann Machines with higher order interactions, and other models.

Lemma 3.A.1. *RBM, DBN, DBM universal approximators of probability distributions on* $\{0,1\}^n$ *contain at least* $2^n - 1$ *parameters.*

Corollary 3.A.2.

- *An RBM universal approximator of distributions on* n *visible units has* $m \geq \frac{2^n - n - 1}{n+1} \sim \frac{2^n}{n}$ *hidden units.*
- *A DBN and a DBM universal approximator on* n *visible units and layers of width* $(n + c)$ *has at least a number of layers of order* $\frac{2^n}{n^2}$ *(and a number of hidden units of order* $\frac{2^n}{n}$*).*

Interestingly, for deep and narrow architectures and for shallow architectures, the bound on the number of hidden units is the same. We make the following informal observation: Minimizing the number of hidden units $\sum_l n_l$ (n_l is the number of units in layer l) of a network with pairwise interactions between neighboring layers while keeping the number of parameters $\sum_l n_{l-1} n_l + \sum_l n_l$ constant to the minimal necessary value $(2^n - 1)$ yields something of the form: Two hidden layers of size $\sqrt{2^n}$ (and a total of $2\sqrt{2^n}$ hidden units).

3.B A Test of Universal Approximation

When represented as mixtures of products, the probability distributions supported by the perfect binary codes of minimum distance two require the maximal number of mixture components (see Corollary 1.3.3). Therefore, it is appealing to use u_{Z_+} to test whether a binary stochastic network is a universal approximator. In this short appendix we test the representability of such probability distributions by RBMs of different sizes. In Chapter 4 we will show that RBMs can approximate $u_{Z_{\pm,n}}$ better than mixtures of independence models with the same dimension, and we will show that $\mathrm{RBM}_{n,n-1}$ can't approximate $u_{Z_{\pm,n}}$ when n is odd and larger than one.

The visible distribution of an RBM with parameters W, B, C is $p = \frac{1}{Z}\sum_h \exp(hWv + Bv + Ch)$. If p has support Z_+, then for every h the function $\exp(hWv + Bv + Ch)$ must be proportional to a point measure (see Corollary 1.3.3), i.e., there must exist a $v' \in \{0,1\}^n$ such that $p(v'|h)/p(v|h) = \infty$ for all $v \neq v'$. This can be expressed as $(hW + B)(v' - v) = \infty$ for all $v \neq v'$, and

$$(\underline{h}|\mathbb{1}^\top)\begin{pmatrix} W & \underline{C} \\ \underline{B} & 0 \end{pmatrix}\begin{pmatrix} \underline{v} \\ \mathbb{1} \end{pmatrix} \stackrel{!}{=} \begin{pmatrix} -\infty & -\infty & \cdots \\ \lambda^{1,1} & -\infty & \cdots \\ -\infty & \lambda^{2,1} & -\infty \\ \lambda^{1,k_1} & -\infty & \cdots \end{pmatrix} =: \underline{\lambda}\,, \tag{3.12}$$

where $\underline{h} \in \{0,1\}^{2^m \times m}$ and $\underline{v} \in \{0,1\}^{n \times 2^n}$ are matrices with the hidden and visible states as rows, and $\mathbb{1} = (1, \ldots, 1)$ is a row vector of appropriate length. The matrix in the right hand side has 2^n columns, each one corresponding to a visible state vector, and 2^m rows, each one corresponding to a hidden state vector. In each row only one entry may be different from $-\infty$, and for the values differing from $-\infty$ in the column i, $\sum_l \exp(\lambda^{i,l}) \propto p(v_i)$. The above equation can be reformulated as a linear equation[1]:

$$\left(\begin{pmatrix} \underline{v} \\ \mathbb{1} \end{pmatrix}^\top \otimes (\underline{h}|\mathbb{1}^\top)\right) \operatorname{vec}\begin{pmatrix} W & \underline{C} \\ \underline{B} & 0 \end{pmatrix} = \operatorname{vec}\underline{\lambda}\,, \tag{3.13}$$

where the vector $\operatorname{vec} M$ is the concatenation of all columns of M into a single column, and $\otimes$ denotes the usual Kronecker product. We tested the existence of solutions to eq. (3.13) for the special case where all $\lambda^{i,k}$ are set to 0, and all other values are only required to be different from 0 and have common sign. We treated this linear programming problem with MATLAB and found the following:

(n,m)	$(2,2)$	$(2,3)$	**$(3,2)$**	$(4,2)$	$(3,3)$	$(3,4)$	$(4,3)$	$(4,5)$
parameter bound satisfied (Corollary 3.A.2)	yes	yes	**yes**	no	yes	yes	yes	yes
eq. (3.13) has a solution	yes	yes	**no**	no	yes	yes	yes	yes

Hence $\mathrm{RBM}_{3,2}$ is not a universal approximator, although it has $3 + 2 + 6 = 11$ parameters, which is more than $\dim(\mathcal{P}_3) = 7$. In Section 4.B we will discuss this interesting model in greater detail, and give an analytical proof of the above statement. By Theorem 3.1.1, $\mathrm{RBM}_{3,3}$ is the smallest RBM universal approximator on $\{0,1\}^3$. It has 15 parameters, which is more than twice the dimension of $\mathcal{P}_3$.

[1] The equivalence $AXB = C \Leftrightarrow (B^\top \otimes A)\operatorname{vec}(X) = \operatorname{vec}(C)$ is sometimes called *Roth's column lemma*, see [98].

3.C A Numerical Comparison

It is known that deep neural networks can represent certain functions way more compactly than shallow networks (see, e.g., [17, 62]). It is reasonable that RBMs and DBNs involving the same number of parameters represent different subsets of the probability simplex and that these subsets are not contained in each other, unless one of the models is a universal approximator. It is generally expected that deeper systems can represent more "complicated" functions than shallow systems, and that deep systems are better at approximating "interesting" probability distributions. It is not easy to prove or disprove the correctness of this intuition. In the first place, this demands a formal definition of "complicated" and "interesting". The currently most accepted picture is that there is a "right" deepness for each particular class of problems under consideration. A conclusive assessment of this question would signify an important advance in the field. The present appendix intends to give an informal, intuitive picture of the maps $\mathbb{R}^d \to \mathrm{RBM}_{n,m}$ and $\mathbb{R}^d \to \mathrm{DBN}_{n_0,n_1,\ldots,n_l}$. To this end we visualize one-dimensional submodels and probability distributions sampled at random from RBMs and DBNs. In Chapter 4 we will pursue a formal analysis of the classes of probability distributions that can be represented by RBMs and DBNs, as well as their approximation errors.

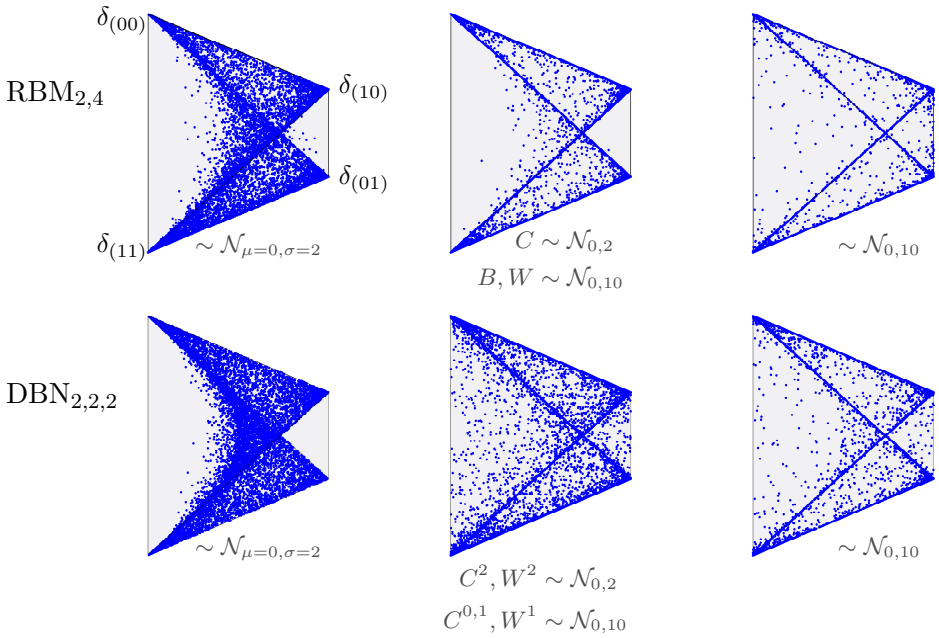

Figure 3.3: *Ten thousand probability distributions on* $\{0,1\}^2$ *sampled at random from the universal approximators* $\mathrm{RBM}_{2,4}$ *and* $\mathrm{DBN}_{2,2,2}$.

In Figure 3.3 we compare the models $\mathrm{RBM}_{2,4}$ and $\mathrm{DBN}_{2,2,2}$. Both models have 2 visible units, 4 hidden units, and the same number of parameters, 12. Both models are excessively overparameterized universal approximators ($\mathrm{RBM}_{2,1}$ is a universal approximator contained in both models, see Theorem 3.1.1). We sampled ten thousand probability distributions from each model (the top row of Figure 3.3 shows the RBM and the bottom row shows the DBN) using three different priors on their standard parameter spaces (each column corresponds to a different prior). The RBM parameters are the bias weights $B \in \mathbb{R}^2$ of the visible units, the bias weights

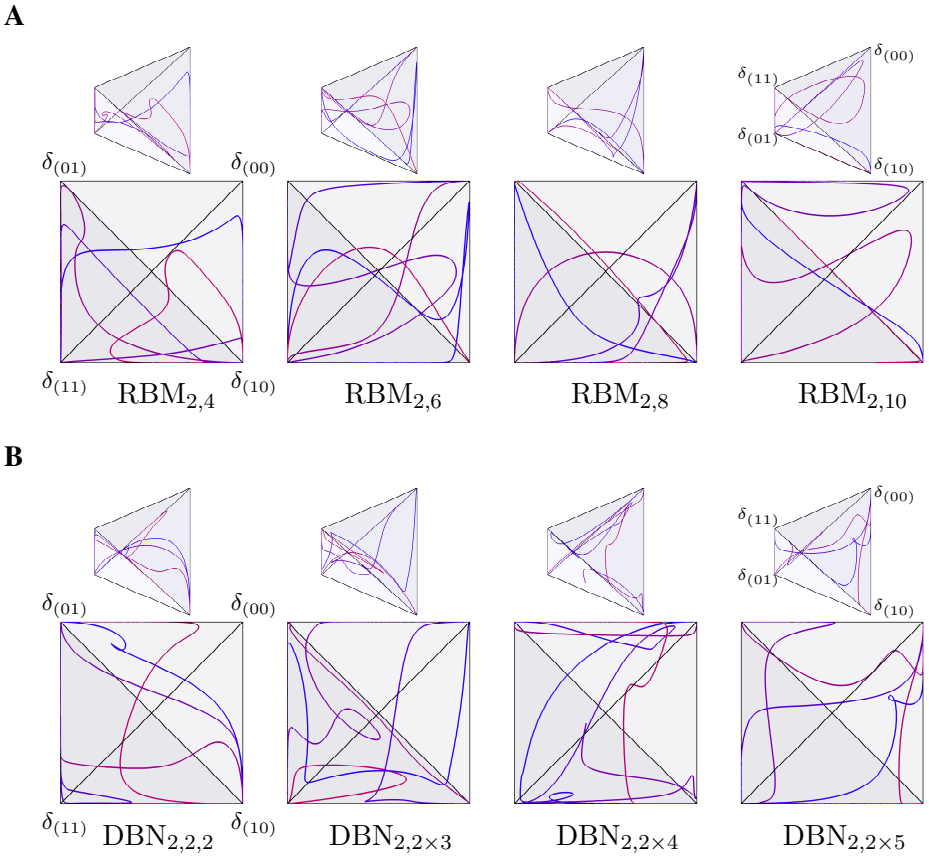

Figure 3.4: *Random one-dimensional submodels of RBMs and DBNs with two visible units.*

$C \in \mathbb{R}^4$ of the hidden units, and the connection weights $W \in \mathbb{R}^{2\times3} \simeq \mathbb{R}^6$. The DBN parameters are the bias weights $C^0, C^1, C^2 \in \mathbb{R}^2$ of the visible nodes, first hidden layer, and second hidden layer respectively, and the connection weights $W^1, W^2 \in \mathbb{R}^{2\times2} \simeq \mathbb{R}^4$. $\mathcal{N}_{\mu,\sigma}$ denotes the multivariate normal distribution with mean μ and variance σ^2. The figures show that the dispersion of the probability distributions is different for the two models. The probability distributions with weights concentrated around 0 and standard deviation 2 lie close to the independence model (left figures). See Figure 1.1 left for the independence model on $\{0,1\}^2$. If the variance of B is too small, then the sampled distributions concentrate around the uniform distribution (not shown). If the variance is too large, the sampled distributions eventually concentrate at the boundary of the independence model (right figures). Since both models are universal approximators, there is a prior on the parameter space for which the sampled probability distributions are uniformly distributed on the visible probability simplex (w.r.t., e.g., the Lebesgue measure or Jeffreys prior). In this experiment we found a simple approximation of such a prior for the DBN, but not for the RBM (middle). Sampling the parameters of the DBN from a normal distribution produces a fairly homogeneous distribution of the visible probability distributions in the probability simplex. In the case of the RBM, the visible probability distributions tend to concentrate at some regions of the probability simplex.

Figure 3.4 A: Each column shows the probability simplex $\mathcal{P}_2$ and five one-dimensional submodels of the RBM with two visible units and 4 to 10 hidden units. The submodels correspond

to one-dimensional randomly chosen affine subspaces of the usual parameter space of the RBMs. Let d be the number of parameters of the RBM. We chose a direction vector r uniformly at random in the sphere S^{d-1} (i.e., $\hat{r} = \frac{r}{\|r\|_2}, r_i \sim \mathcal{N}_{0,1}$) and an origin vector $s \in \mathbb{R}^d, s_i \sim \mathcal{N}_{0,2}$. We plotted the probability distributions in $\mathrm{RBM}_{n,m}$ for which the parameters W, B and C satisfy $\mathrm{vec}(W, B, C) = \alpha(r + s)$, and $\alpha \in \mathbb{R}$ (the figures show α with norm ≤ 100; most of the time for $|\alpha| \approx 30$ the resulting probability distributions are very close to the boundary of the probability simplex). These one-dimensional models are marginals of exponential geodesics on $\{0, 1\}^{n+m}$. In contrast to exponential geodesics, these one-dimensional models can have the same limit point for $\alpha \to \infty$ and $\alpha \to \infty$ (see Chapter 2). Figure 3.4 B shows five one-dimensional submodels of DBN models with 2 visible units, computed in a similar way as the submodels of RBMs depicted in Figure 3.4 A.

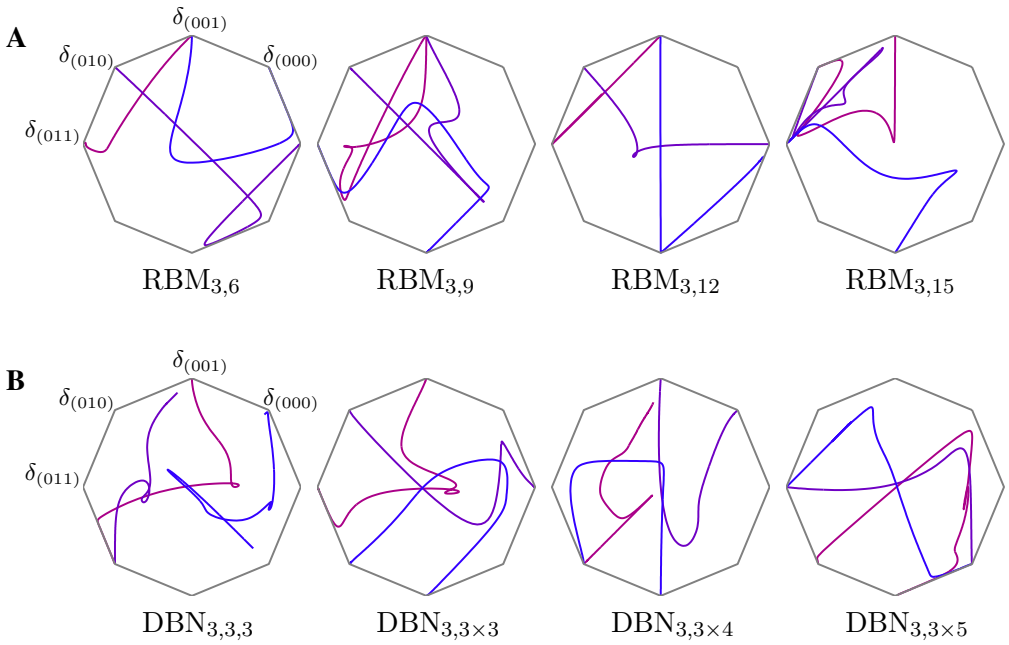

Figure 3.5: *Random one-dimensional submodels of RBMs and DBNs with three visible units.*

Figure 3.5 A: Each figure shows the projection of the probability simplex $\mathcal{P}_3$ onto a polygon with 2^3 vertices. The curves are three one-dimensional submodels of the RBM with 3 visible nodes. These submodels correspond to one-dimensional randomly chosen affine subspaces of the parameter space of the RBMs, as in Figure 3.4. We plotted all probability vectors with parameter vectors of norm less or equal to 100. Figure 3.5 B: Each figure shows three one-dimensional submodels of DBN (as in the previous figures). In this case we plotted all probability distributions with parameter vectors of norm less or equal to 200. For parameter vectors with norm less than 100 the curves were often not very close to the boundary of the polygon. Seemingly, the DBN submodels are more concentrated in the interior of the polygons than those of the RBMs. In such a case they can *(i)* have a derivative with smaller magnitude, *(ii)* be more "curvy", or *(iii)* their limit points are less likely to approach point measures. Either case is interesting. Note that a random exponential geodesic has almost surely limit points which are point measures (see Proposition 2.3.11). The one-dimensional submodels of RBMs are projections of exponential geodesics.

4 Expressive Power and Approximation Errors of RBMs and DBNs

In this chapter we present explicit classes of probability distributions that can be learned by Restricted Boltzmann Machines and Deep Belief Networks depending on the number of hidden units and hidden layers that they contain. Using these descriptions, we estimate the maximal Kullback-Leibler divergence between arbitrary probability distributions in the probability simplex and their best approximations within the RBM and DBN models. We show that the maximal KL-divergence to the RBM model with n visible and m hidden units is bounded from above roughly by $(n-1) - \log(m+1)$, and we show analogue results for DBNs. In this way we can control the number of hidden units and hidden layers that guarantee sufficiently rich models respecting a given error tolerance.

This chapter begins with a brief discussion of different kinds of errors. Section 4.1 introduces and studies special mixtures of independence models and partition models, the statistical models that we use to assess RBMs and DBNs. Section 4.2 contains the main results on the expressive power and approximation errors of RBMs. Section 4.3 contains the main results on the expressive power the approximation errors of DBNs. In Appendix 4.A we show that the smallest mixture model of independence models which contains an RBM model is very large. Appendix 4.B contains an elaboration on the models $\mathrm{RBM}_{3,2}$ and $\mathrm{RBM}_{4,2}$.

Preliminaries

As shown in Section 3, the model $\mathrm{RBM}_{n,m}$ with $m \geq 2^{n-1}-1$, and the DBN model with $2^n/2n$ layers of width n are universal approximators. An RBM or a DBN with layers of equal width which are universal approximators of distributions on $\{0,1\}^n$ have at least $\dim(\mathcal{P}) = 2^n - 1$ parameters and $\lceil 2^n/(n+1) \rceil - 1$ hidden units. The geometry of $\mathrm{RBM}_{n,m}$ is intricate, and even an RBM of dimension $2^n - 1$ is not guaranteed to contain all visible distributions. In Section 3.B we showed that $\mathrm{RBM}_{3,2}$ is an example of this behavior.

In practice, training such large systems is not desirable or even possible. This was already pointed out in Freund and Haussler's seminal paper on RBMs [45]. There are at least two reasons why in many cases it is not necessary to have universal approximators:

- An appropriate approximation of distributions is sufficient for most purposes.
- The interesting distributions that the system shall simulate belong to a small class of distributions.

For example, the set of optimal policies in reinforcement learning [111], the set of dynamics kernels that maximize predictive information in robotics [121], or the information flow in neural networks [15] are contained in very low dimensional manifolds (see Section 5.2). On the other hand, usually it is very hard to mathematically describe sets containing the optimal solutions to general problems, or sets of interesting probability distributions (for example the set of

distributions generating natural images). Furthermore, although RBMs and DBNs are parametric models and in theory for any choice of the parameters it is possible to compute the resulting probability distribution, in practice it is difficult to explicitly specify this probability distribution, or even to estimate it (see [76] for an account on RBMs). Due to these difficulties the number of hidden units and hidden layers is often chosen on the basis of experience [57], or is considered as a hyperparameter which is optimized by extensive search.

Approximation Error

When training a system to represent a distribution p, there are mainly three contributions to the discrepancy between p and the state of the system after training:

- Usually, the underlying distribution p is unknown and only a set of samples generated by p is observed. These samples can be represented as an empirical distribution p^{Data}, which usually is not identical with p.
- The set $\mathrm{RBM}_{n,m}$, respectively $\mathrm{DBN}(n_0^l)$ does not contain every probability distribution, unless the number of hidden units is very large, as we outlined above. Therefore, we have an approximation error given by the distance of p^{Data} to its best approximation $p^{\text{Data}}_{\text{Model}}$ within the model.
- The learning process may yield a solution $\tilde{p}^{\text{Data}}_{\text{Model}}$ within the model, which is not the optimum $p^{\text{Data}}_{\text{Model}}$. This occurs, for example, if the learning algorithm gets trapped in a local optimum, or if it optimizes an objective different from *maximum likelihood*, e.g., contrastive divergence (CD), see [25].

Le Roux and Bengio [72] show that the log-likelihood of a target distribution can be strictly improved by increasing the number of hidden units of an RBM unless the RBM is a universal approximator (see also [17, Chapter 5.3]). We are interested in quantifying the expressive power of the RBM and DBN models and the Kullback-Leibler divergence from an arbitrary distribution to its best representation within the models. Estimating the approximation error is difficult, because the geometry of these models is not sufficiently understood. Our strategy is to find subsets $\mathcal{M} \subseteq \mathrm{RBM}_{n,m}$, $\mathcal{M}' \subset \mathrm{DBN}(n_0^l)$ that are relatively easy to describe. The maximal error when approximating probability distributions with the model is upper bounded by the maximal error when approximating with the submodels.

Kullback-Leibler Divergence and Reversed Information Projections

If $\mathcal{E} \subseteq \mathcal{P}$ is a statistical model and $p \in \mathcal{P}$, then any probability distribution $p_{\mathcal{E}} \in \overline{\mathcal{E}}$ satisfying

$$D(p\|p_{\mathcal{E}}) = D(p\|\mathcal{E}) := \min\{D(p\|q) : q \in \overline{\mathcal{E}}\}$$

is called a *(generalized) reversed information projection*, or rI-projection. If p is an empirical distribution, then one can show that any rI-projection is a maximum likelihood estimate. Exponential families behave nicely with respect to rI-projections. If $\mathcal{E}$ is an exponential family, any $p \in \mathcal{P}$ has a unique rI-projection $p_{\mathcal{E}}$ to $\mathcal{E}$, see [10].

In order to assess some model $\mathcal{M}$ we use the maximal approximation error with respect to the KL-divergence when approximating arbitrary probability distributions using $\mathcal{M}$:

$$D_{\mathcal{M}} := \sup\{D(p\|\mathcal{M}) : p \in \mathcal{P}\} \ .$$

For example, if $\mathcal{M}$ consists only of the uniform distribution, the maximal KL-divergence is attained by any Dirac delta distribution δ_x, $x \in \mathcal{X}$, and amounts to:

$$D_{\{u\}} = D(\delta_x \| u) = \log |\mathcal{X}| \, . \tag{4.1}$$

Lemma 4.0.1. (Corollary 4.1 of [13]). *Let $n \in \mathbb{N}$ and let $\overline{\mathcal{E}_n^1}$ be the independence model on $\{0,1\}^n$. Then $D_{\mathcal{E}_n^1} = (n-1)$. The global KL-divergence maximizers are the distributions of the form $\frac{1}{2}(\delta_x + \delta_y)$, where $x, y \in \{0,1\}^n$ satisfy $x_i + y_i = 1$ for all $i \in [n]$.*

This result should be compared with eq. (4.1). Although the independence model is much larger than the set $\{u\}$, the maximal divergence decreases only by 1. As shown in [95], if $\mathcal{E}$ is any exponential family of dimension k, then $D_{\mathcal{E}} \geq \log(|\mathcal{X}|/(k+1))$. Thus this notion of distance is rather strong. The exponential families satisfying $D_{\mathcal{E}} = \log(|\mathcal{X}|/(k+1))$ are *partition models*.

4.1 Partition Models and Restricted Mixture Models

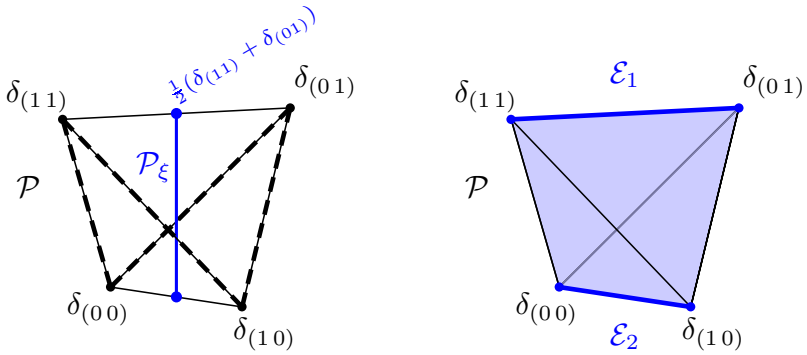

Figure 4.1: *Left: The blue line represents the partition model $\mathcal{P}_\xi$ with partition $\xi = \{\{(11),(01)\},\{(00),(10)\}\}$. The dashed lines represent the set of KL-divergence maximizers for $\mathcal{P}_\xi$. Right: The mixture of the product distributions $\mathcal{E}_1$ and $\mathcal{E}_2$ with disjoint supports on $\{(11),(01)\}$ and $\{(00),(10)\}$ equals the whole probability simplex.*

The mixture $\mathrm{Mixt}(\mathcal{M}_1, \ldots, \mathcal{M}_m)$ of m models $\mathcal{M}_1, \ldots, \mathcal{M}_m \subseteq \mathcal{P}$ is the set of all convex combinations

$$q = \sum_{i \in [m]} \alpha_i q_i \, , \text{ where } q_i \in \mathcal{M}_i, \alpha_i \geq 0, \ \forall i \in [m] \text{ and } \sum_{i \in [m]} \alpha_i = 1 \, . \tag{4.2}$$

In general mixture models are complicated objects. Even if all models $\mathcal{M}_1 = \cdots = \mathcal{M}_m$ are equal, it is difficult to describe the mixture (see Chapter 1). The situation simplifies considerably if the models have disjoint supports. Given any partition $\xi = \{\mathcal{X}_1, \ldots, \mathcal{X}_m\}$ of $\mathcal{X}$, any $p \in \mathcal{P}$ can be written as $p(x) = p(\mathcal{X}_i)p(x|\mathcal{X}_i)$ for all $x \in \mathcal{X}_i$ and $i \in \{1, \ldots, m\}$, where $p(\cdot|\mathcal{X}_i)$ is a probability measure in $\mathcal{P}(\mathcal{X}_i)$ for all i.

Lemma 4.1.1. *Let $\xi = \{\mathcal{X}_1, \ldots, \mathcal{X}_m\}$ be a partition of $\mathcal{X}$ and let $\mathcal{M}_1, \ldots, \mathcal{M}_m$ be statistical models such that $\mathcal{M}_i \subseteq \mathcal{P}(\mathcal{X}_i)$ for $i \in [m]$. Consider any $p \in \mathcal{P}$ and let p_i be an*

rI-projection of $p(\cdot|\mathcal{X}_i)$ to $\mathcal{M}_i$ for $i \in [m]$. Then the rI-projection $p_\mathcal{M}$ of p to the mixture $\mathcal{M} = \mathrm{Mixt}(\mathcal{M}_1, \ldots, \mathcal{M}_m)$ satisfies

$$p_\mathcal{M}(x) = p(\mathcal{X}_i)p_i(x), \qquad \forall x \in \mathcal{X}_i \; \forall i \in [m] \; .$$

Therefore, $D(p\|\mathcal{M}) = \sum_i p(\mathcal{X}_i)D(p(\cdot|\mathcal{X}_i)\|\mathcal{M}_i)$, and $D_\mathcal{M} = \max_{i=1,\ldots,m} D_{\mathcal{M}_i}$.

Proof. Let $q \in \mathcal{M}$ be as in eq. (4.2). Then

$$D(p\|q) = \sum_{i=1}^{m} p(\mathcal{X}_i)D(p(\cdot|\mathcal{X}_i)\|q_i) + \sum_{i=1}^{m} p(\mathcal{X}_i)\log\frac{p(\mathcal{X}_i)}{\alpha_i} \tag{4.3}$$

for all $p \in \mathcal{P}$. For fixed p the first sum is minimal if and only if each term is minimal. The second sum vanishes for $\alpha_i = p(\mathcal{X}_i)$. □

Remark 4.1.2. If each $\mathcal{M}_i$ is an exponential family on $\mathcal{X}_i$, then the mixture $\mathrm{Mixt}(\mathcal{M}_1, \ldots, \mathcal{M}_m)$ is an exponential family of dimension $\dim(\mathrm{Mixt}(\mathcal{M}_1, \ldots, \mathcal{M}_m)) = \sum_{i\in[m]}(\dim(\mathcal{M}_i)+1)-1$ (this is not true if the supports of the models $\mathcal{M}_i$ are not disjoint). If A^i is a sufficient statistics of $\mathcal{M}_i$, one may assume that A^i contains the row $\mathbb{1}_{\mathcal{X}_i}$. The block diagonal

$$\oplus_{i\in[m]}A^i = \begin{pmatrix} A^i & & \\ & \ddots & \\ & & A^m \end{pmatrix}$$

is then a sufficient statistics of the mixture $\mathrm{Mixt}(\mathcal{M}_1, \ldots, \mathcal{M}_m)$.

Definition 4.1.3. If ξ is a partition of $\mathcal{X}$ with blocks $\{\mathcal{X}_i\}_{i\in[m]}$ and $\mathcal{M}_i$ equals the set containing just the uniform distribution on $\mathcal{X}_i$ for all $i \in [m]$, then $\mathrm{Mixt}(\mathcal{M}_1, \ldots, \mathcal{M}_m)$ is called the *partition model* with partition ξ, denoted with $\mathcal{P}_\xi$.

The partition model $\mathcal{P}_\xi$ consists of all distributions with constant value on each block $\mathcal{X}_i$, i.e., those with $p(x) = p(y)$ for all $x, y \in \mathcal{X}_i$ for all i. This is the closure of the exponential family with sufficient statistics

$$A_x = (\mathbb{1}_{\mathcal{X}_1}(x), \mathbb{1}_{\mathcal{X}_2}(x), \ldots, \mathbb{1}_{\mathcal{X}_d}(x))^\top \; .$$

A partition model is a convex exponential family with uniform reference measure (see Chapter 2). The partition models include the set of finite exchangeable distributions, where the blocks of the partition are the sets of binary vectors which have the same number of entries equal to one. The probability of a vector v depends only on the number of ones, but not on their position. See [37] for interesting properties of exchangeable distributions. See Figure 4.1 for a small example of a partition model. A key property of partition models is that they minimize $D_\mathcal{E}$ among all exponential families of a fixed dimension. See [95] for interesting properties of partition models.

As a consequence of Lemma 4.1.1 and eq. (4.1) we have:

Corollary 4.1.4. *Let $\xi = \{\mathcal{X}_1, \ldots, \mathcal{X}_m\}$ be a partition of $\mathcal{X}$. Then $D_{\mathcal{P}_\xi} = \max_{i=1,\ldots,m} \log|\mathcal{X}_i|$.*

See Figure 4.2 for an intuition on the approximation error of partition models depending on the block sizes.

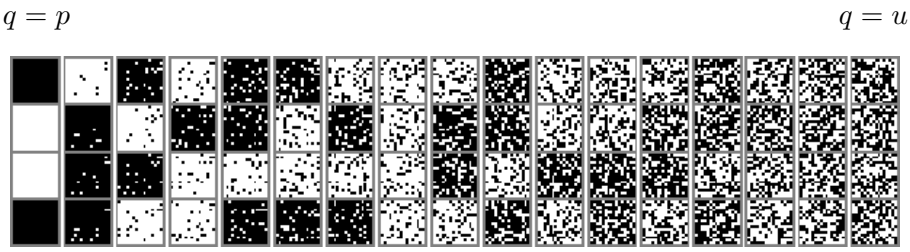

Figure 4.2: *This figure gives an intuition on what the size of an error means for probability distributions on images with* 16×16 *pixels. Every column shows four samples drawn from the best approximation* q *of the distribution* $p = \frac{1}{2}(\delta_{(1\ldots1)} + \delta_{(0\ldots0)})$ *within a partition model with 2 randomly chosen cubical blocks containing* $(0\ldots0)$ *and* $(1\ldots1)$*, of cardinality from* 1 *(first column) to* $\frac{|\mathcal{X}|}{2}$ *(last column). As a measure of error ranging from* 0 *to* 1 *we take* $D(p\|q)/D(p\|u)$*. The last column shows samples from the uniform distribution, which is in particular the best approximation of* p *within* $\mathrm{RBM}_{n,0}$*. Note that* $\mathrm{RBM}_{n,1}$ *can approximate* p *with arbitrary accuracy, see Theorem 3.1.1.*

Now, assume that $\mathcal{X} = \{0,1\}^n$. The vertices of a k-dimensional face of the n-cube (a cubical set) are given by fixing the values of x in $n-k$ positions: $\{x \in \{0,1\}^n : x_i = \tilde{x}_i, \forall i \in I, \text{ for some } I \subseteq \{1,\ldots,n\}, |I| = n-k\}$. A cubical subset of cardinality 2^k can be naturally identified with $\{0,1\}^k$. This identification allows us to define independence models and product measures on $\mathcal{P}(\mathcal{Y}) \subseteq \overline{\mathcal{P}}(\mathcal{X})$. Note that product measures on $\mathcal{Y}$ are also product measures on $\mathcal{X}$, and the independence model on $\mathcal{Y}$ is a subset of the independence model on $\mathcal{X}$. Figure 4.1 shows a small example of a mixture of independence models with disjoint supports.

Corollary 4.1.5. *Let* $\xi = \{\mathcal{X}_1,\ldots,\mathcal{X}_m\}$ *be a partition of* $\mathcal{X} = \{0,1\}^n$ *into cubical sets. For any* i *let* $\mathcal{E}_i$ *be the independence model on* $\mathcal{X}_i$*, and let* $\mathcal{M}$ *be the mixture of* $\mathcal{E}_1,\ldots,\mathcal{E}_m$*. Then*

$$D_{\mathcal{M}} = \max_{i=1,\ldots,m} \log(|\mathcal{X}_i|) - 1 \ .$$

The following lemma will be used in the proof of Theorem 4.2.2:

Lemma 4.1.6. *Let* $n_1,\ldots,n_m$ *be non-negative integers satisfying* $2^{n_1} + \cdots + 2^{n_m} = 2^n$*. Let* $\mathcal{M}$ *be the union of all mixtures of independence models* $\mathcal{E}_i \subseteq \mathcal{P}(\mathcal{X}_i)$ $i \in [m]$ *corresponding to all cubical partitions of* $\mathcal{X}$ *into blocks* $\{\mathcal{X}_i\}_{i\in[m]}$ *of cardinalities* $2^{n_1},\ldots,2^{n_m}$*. Then* $D_{\mathcal{M}} \leq \sum_{i:n_i>1} \frac{n_i-1}{2^{n-n_i}}$.

Proof. The proof is by induction on n. If $n = 1$, then $m = 1$ or $m = 2$, and in both cases it is easy to see that the inequality holds (both sides vanish). If $n > 1$, then order the n_i such that $n_1 \geq n_2 \geq \cdots \geq n_m \geq 0$. Without loss of generality assume $m > 1$.

Let $p \in \mathcal{P}(\mathcal{X})$, and let $\mathcal{Y}$ be a cubical subset of $\mathcal{X}$ of cardinality 2^{n-1} such that $p(\mathcal{Y}) \leq \frac{1}{2}$. Since the numbers $2^{n_1} + \cdots + 2^{n_i}$ for $i = 1,\ldots,m$ contain all multiples of 2^{n_1} up to 2^n, and $2^n/2^{n_1}$ is even, there exists k such that $2^{n_1} + \cdots + 2^{n_k} = 2^{n-1} = 2^{n_{k+1}} + \cdots + 2^{n_m}$.

Let $\mathcal{M}'$ be the union of all mixtures of independence models corresponding to all cubical partitions $\xi = \{\mathcal{X}_1,\ldots,\mathcal{X}_m\}$ of $\mathcal{X}$ into m blocks of cardinalities $n_1,\ldots,n_m$ such that $\mathcal{X}_1 \cup \cdots \cup \mathcal{X}_k = \mathcal{Y}$. In the following, the symbol $\sum_i'$ shall denote summation over all indices i such

that $n_i > 1$. By induction

$$D(p\|\mathcal{M}) \leq D(p\|\mathcal{M}') \leq p(\mathcal{Y}) \sum_{i=1}^{k}{}' \frac{n_i - 1}{2^{n-1-n_i}} + p(\mathcal{X} \setminus \mathcal{Y}) \sum_{j=k+1}^{m}{}' \frac{n_j - 1}{2^{n-1-n_j}} \,. \tag{4.4}$$

There exist $j_1 = k+1 < j_2 < \cdots < j_k < j_{k+1} = m+1$ such that $2^{n_i} = 2^{n_{j_i}} + \cdots + 2^{n_{j_{i+1}-1}}$ for all $i \leq k$. Note that

$$\sum_{j=j_i}^{j_{i+1}-1}{}' \frac{n_j - 1}{2^{n-1-n_j}} \leq \frac{n_i - 1}{2^{n-1}}(2^{n_{j_i}} + \cdots + 2^{n_{j_{i+1}-1}}) = \frac{n_i - 1}{2^{n-1-n_i}} \,,$$

and therefore

$$(\tfrac{1}{2} - p(\mathcal{Y}))\frac{n_i - 1}{2^{n-1-n_i}} \; + \; (\tfrac{1}{2} - p(\mathcal{X} \setminus \mathcal{Y})) \sum_{j=j_i}^{j_{i+1}-1}{}' \frac{n_j - 1}{2^{n-1-n_j}} \geq 0 \,.$$

Adding these terms for $i = 1, \ldots, k$ to the right hand side of eq. (4.4) yields

$$D(p\|\mathcal{M}) \leq \frac{1}{2} \sum_{i=1}^{k}{}' \frac{n_i - 1}{2^{n-1-n_i}} + \frac{1}{2} \sum_{j=k+1}^{m}{}' \frac{n_j - 1}{2^{n-1-n_j}} \,,$$

from which the assertion follows. □

We can also derive lower bounds for the maximal approximation errors of mixtures:

Lemma 4.1.7. *Let $\mathcal{M}$ be the union of all mixtures of m independence models with disjoint supports on $\mathcal{X} = \{0,1\}^n$ (not necessarily partitioning $\mathcal{X}$). Let $\widetilde{\mathcal{M}}$ be the union of all partition models with m cubical blocks. Let C be a binary code of length n, cardinality $m+1$, and minimum distance d. Then*

$$D(u_C\|\mathrm{Mixt}^m(\overline{\mathcal{E}_n^1})) = D(u_C\|\mathcal{M}) = D(u_C\|\widetilde{\mathcal{M}}) = \frac{2(d-1)}{(m+1)} \,.$$

Furthermore, $D(p\|\mathrm{Mixt}^m(\overline{\mathcal{E}_n^1})) = D(p\|\mathcal{M})$ for any p with $\mathrm{supp}(p) = C$.

Remark 4.1.8.

(i) By Lemma 4.1.7 $\frac{2(d-1)}{(m+1)}$ is a lower bound for $D_{\widetilde{\mathcal{M}}}$, $D_{\mathcal{M}}$, and $D_{\mathrm{Mixt}^m(\overline{\mathcal{E}_n^1})}$.

(ii) If $m = 1$ and C is a binary code consisting of two antipodal points, i.e., $C = \{x, \mathbb{1} - x\}$, then $d = n$. The lower and upper bounds on the approximation errors given in Lemma 4.1.7 and Lemma 4.1.6 yield $D_{\mathcal{E}_n^1} = n - 1$. This resembles the previous results on the approximation errors of independence models given in Lemma 4.0.1.

Proof of Lemma 4.1.7. (i) Consider an arbitrary family of disjoint cubical blocks $\{\mathcal{X}_i\}_{i\in[m]}$. We have

$$\begin{aligned} D(u_C\|\mathrm{Mixt}(\mathcal{E}_1, \ldots, \mathcal{E}_m)) &= \sum_{x\in C} \frac{1}{|C|} \log\left(\frac{1/|C|}{p}\right) \\ &= \frac{1}{|C|} \sum_{i\in[m]} \sum_{x\in\mathcal{X}_i\cap C} \log\left(\frac{1/|C|}{\lambda_i/|\mathcal{X}_i|}\right) = \log 1/|C| - \sum_{i\in[m]} \frac{|\mathcal{X}_i \cap C|}{|C|} \log(\lambda_i/|\mathcal{X}_i|) \,. \end{aligned}$$

Here we set $p(x) = \lambda_i \frac{1}{|\mathcal{X}_i|}$ for all $x \in \mathcal{X}_i$, which is justified by Lemma 4.1.1. The expression $\sum_{i\in[m]} \frac{|\mathcal{X}_i \cap C|}{|C|}(\log(|\mathcal{X}_i|) - \log(\lambda_i))$ is minimized for $\lambda_i = \frac{|\mathcal{X}_i \cap C|}{|C|}$ (this follows using Lagrange multipliers), and if $|C| = m+1$, the expression $\log(1/|C|) + \sum_{i\in[m]} \frac{|\mathcal{X}_i \cap C|}{|C|}(\log(|\mathcal{X}_i|) - \log(\frac{|\mathcal{X}_i \cap C|}{|C|}))$ is minimized for $|\mathcal{X}_i| = 1$, for $i = 1, \ldots, m-1$, and $|\mathcal{X}_m| = 2^d$, where d is the minimal distance of C, and takes the claimed value.

(ii) Let r have support C, and q be an rI-projection of r to $\mathrm{Mixt}^m(\overline{\mathcal{E}_n^1})$. The divergence $D(r\|q)$ depends on $q(\cdot|C)$ and $q(C)$. For any choice of $q(\cdot|C)$ increasing $q(C)$ reduces the divergence. Any $q(\cdot|C)$ can be realized as a mixture of products with disjoint supports. The value of $q(C)$ is maximized by mixture whose components have disjoint supports. □

Proposition 4.1.9. *Let $\mathcal{M}$ be the union of all mixtures of m independence models with disjoint supports on $\mathcal{X} = \{0,1\}^n$ (not necessarily partitioning $\mathcal{X}$). Then*

$$D(u_{Z_{\pm,n}}\|\mathcal{M}) = 1 - \frac{K}{2^{n-1}}\ , \tag{4.5}$$

where K is the maximal number of blocks of cardinality two in a cubical partition of $\mathcal{X}$ with m blocks. In particular, $K = 0$ for $m < n$, $K = 2$ for $m = n$, $K = 4$ for $m = n+1$, and $K = 2^{n-1}$ only for $m \geq 2^{n-1}$.

Proof. Let $\sum_{i=1}^m \alpha_i q_i$ be an rI-projection of $u_{Z_{\pm,n}}$ to $\mathcal{M}$. This is the rI-projection to some $\mathrm{Mixt}(\mathcal{E}_{\mathcal{X}_1}, \ldots, \mathcal{E}_{\mathcal{X}_m})$. We may assume that a collection of disjoint cubical blocks $\{\mathcal{X}_i\}$ which covers $Z_{\pm,n}$ is in fact a partition of $\{0,1\}^n$. If a block $\mathcal{X}_i$ has cardinality two, we may choose a point measure for q_i in $\mathcal{E}_{\mathcal{X}_i}$. We write $u_{Z_{\pm,n}}(x) = \frac{|\mathcal{X}_i|}{2^n}\frac{1}{|Z_{\pm,n}\cap\mathcal{X}_i|}$ for any x in the block $\mathcal{X}_i$. As in eq. (4.3), the divergence is given by

$$D(u_Z\|q) = \sum_{i=1}^m \frac{|\mathcal{X}_i|}{2^n} D(u_Z(\cdot|\mathcal{X}_i)\|q_i) + \sum_{i=1}^m \frac{|\mathcal{X}_i|}{2^n}\log\frac{|\mathcal{X}_i|/2^n}{\alpha_i}$$

The second sum vanishes for an appropriate choice of α. The term $D(u_Z(\cdot|\mathcal{X}_i)\|q_i)$ is 1 whenever $|\mathcal{X}_i| > 2$ and vanishes for $|\mathcal{X}_i| = 2$. Hence the total divergence is 1 minus the number of times that a block has cardinality two. □

Example 4.1.10. Let $\mathcal{X} = \{0,1\}^3$ and let $\mathcal{M}$ be the union of all mixtures of three independence models with disjoint supports. The set $\mathcal{X}$ can be partitioned into one face of dimension two and two faces of dimension one. The binary code $Z_{\pm,3}$ has cardinality four and minimum distance two. Lemma 4.1.6 and Lemma 4.1.7 yield

$$D(u_{Z_\pm}\|\mathrm{Mixt}^3(\overline{\mathcal{E}_3^1})) = D_{\mathrm{Mixt}^3(\overline{\mathcal{E}_3^1})} = D(u_{Z_\pm}\|\mathcal{M}) = D_{\mathcal{M}} = \frac{1}{2}\ . \tag{4.6}$$

Hence $u_{Z_\pm}$ are the global maximizers of KL-divergence for $\mathrm{Mixt}^3(\overline{\mathcal{E}_3^1})$ and $\mathcal{M}$.

4.2 Restricted Boltzmann Machines

Consider a set $\xi = \{\mathcal{X}_i\}_{i=1}^m$ of m disjoint cubical subsets of $\mathcal{X}$ (not necessarily partitioning $\mathcal{X}$). We write G_m for the collection of all such sets of sets. We have the following:

Theorem 4.2.1. $\mathrm{RBM}_{n,m}$ *contains the following distributions:*

- *Any mixture of one arbitrary product distribution, k product distributions with support on arbitrary but disjoint faces of the n-cube, and $(m-k)$ arbitrary distributions with support on any edges of the n-cube, for any $0 \leq k \leq m$. In particular:*
- *Any mixture of $(m+1)$ product distributions with disjoint supports. In consequence, $\mathrm{RBM}_{n,m}$ also contains the partition model of any partition in G_{m+1}.*

Restricting the supports of the second item to pairs of vectors differing in one entry shows that an RBM with $m \geq 2^{n-1} - 1$ hidden units is a universal approximator on $\{0,1\}^n$. Hence Theorem 4.2.1 contains Theorem 3.1.1 from the previous chapter as a special case.

Assume $m+1 = 2^k$ and let ξ be a partition of $\mathcal{X}$ into $(m+1)$ disjoint cubical sets of equal size. Denote $\mathcal{P}_{\xi,1}$ the set of all distributions which can be written as a mixture of $(m+1)$ product distributions with support on the blocks of ξ. The dimension of $\mathcal{P}_{\xi,1}$ is given by

$$\dim \mathcal{P}_{\xi,1} = (m+1)\log\left(\frac{2^n}{m+1}\right) + m = (m+1)\cdot n + m - (m+1)\log(m+1)\ .$$

The dimension of the set of visible distributions represented by an RBM is at most equal to the number of parameters $m\cdot n+m+n$. Since $\dim \mathcal{P}_{\xi,1} - \dim \mathrm{RBM}_{m-1} \geq n+1-(m+1)\log(m+1)$, the set $\mathcal{P}_{\xi,1}$, which by Theorem 4.2.1 can be represented by $\mathrm{RBM}_{n,m}$, is not contained in $\mathrm{RBM}_{n,m-1}$ when $(m+1)^{m+1} \leq 2^{n+1}$.

Proof of Theorem 4.2.1. The proof draws on ideas from Section 3.1. An RBM with no hidden units can represent precisely the independence model, and in particular any uniform distribution on a face of the n-cube.

Consider an RBM with $m-1$ hidden units. For any choice of the parameters $W \in \mathbb{R}^{(m-1)\times n}$, $B \in \mathbb{R}^n$, $C \in \mathbb{R}^{m-1}$ we can write the resulting distribution on the visible units as:

$$p(v) = \frac{\sum_h z(v,h)}{\sum_{v',h'} z(v',h')}\ , \tag{4.7}$$

where $z(v,h) = \exp(hWv + Bv + Ch)$. An additional hidden unit with connection weights w to the visible units and bias c, produces a new distribution which can be written as follows:

$$p_{w,c}(v) = \frac{(1+\exp(wv+c))\sum_h z(v,h)}{\sum_{v',h'}(1+\exp(wv'+c))z(v',h')}\ . \tag{4.8}$$

Consider now any set $I \subseteq [n] := \{1,\ldots,n\}$ and an arbitrary visible vector $u \in \mathcal{X}$. The values of u in the positions $[n] \setminus I$ define a face $F := \{v \in \mathcal{X} : v_i = u_i\ , \forall i \notin I\}$ of the n-cube of dimension $|I|$. Let $\mathbb{1} := (1,\ldots,1) \in \mathbb{R}^n$ and denote by $u^{I,0}$ the vector with entries $u_i^{I,0} = u_i, \forall i \notin I$ and $u_i^{I,0} = 0, \forall i \in I$. Let $\lambda^I \in \mathbb{R}^n$ with $\lambda_i^I = 0\ , \forall i \notin I$ and let $\lambda_c, a \in \mathbb{R}$. Define the connection weights w and c as follows:

$$w = a(u^{I,0} - \frac{1}{2}\mathbb{1}^{I,0}) + \lambda^I\ ,$$
$$c = -a(u^{I,0} - \frac{1}{2}\mathbb{1}^{I,0})^\top u + \lambda_c\ .$$

For this choice and $a \to \infty$ eq. (4.8) yields:

$$p_{w,c}(v) = \begin{cases} \frac{p(v)}{1+\sum_{v' \in F} \exp\left(\lambda^I \cdot v' + \lambda_c\right) p(v')}, & \forall v \notin F \\ \frac{(1+\exp(\lambda^I \cdot v + \lambda_c)) p(v)}{1+\sum_{v' \in F} \exp\left(\lambda^I \cdot v' + \lambda_c\right) p(v')}, & \forall v \in F \end{cases} . \tag{4.9}$$

If the initial p from eq. (4.7) is such that its restriction to F is a product distribution, then $p(v) = K \exp(\eta^I \cdot v)\ , \forall v \in F$, where K is a constant and η^I is a vector with $\eta^I_i = 0\ , \forall i \notin I$. We can choose $\lambda^I = \beta^I - \eta^I$ and $\exp(\lambda_c) = \alpha \frac{1}{K \sum_{v \in F} \exp(\beta^I \cdot v)}$. For this choice, eq. (4.9) yields:

$$p_{w,c} = (\alpha - 1)p + \alpha \hat{p}\ ,$$

where $\hat{p}$ is a product distribution with support in F and arbitrary natural parameters β^I, and α is an arbitrary mixture weight in $[0, 1]$. Finally, the product distributions on edges of the cube are arbitrary, see Chapter 1, and hence the restriction of any p to any edge is a product distribution. □

Maximal Approximation Errors of RBMs

Let $m < 2^{n-1} - 1$. By Theorem 4.2.1 all partition models for partitions of $\{0, 1\}^n$ into $(m + 1)$ cubical sets are contained in $\mathrm{RBM}_{n,m}$. Applying Corollary 4.1.4 to a cubical partition where the cardinality of all blocks is at most $2^{n-\lfloor \log(m+1) \rfloor}$ yields the bound $D_{\mathrm{RBM}_{n,m}} \leq n - \lfloor \log(m+1) \rfloor$. Similarly, using mixtures of product distributions, Theorem 4.2.1 and Corollary 4.1.5 imply the smaller bound $D_{\mathrm{RBM}_{n,m}} \leq n - 1 - \lfloor \log(m + 1) \rfloor$. Using Lemma 4.1.6 we derive an improved bound which strictly decreases, as m increases, until 0 is reached:

Theorem 4.2.2. *Let $m \leq 2^{n-1} - 1$. Then the maximal Kullback-Leibler divergence from any distribution on $\{0, 1\}^n$ to $\mathrm{RBM}_{n,m}$ is upper bounded by*

$$\max_{p \in \mathcal{P}} D(p \| \mathrm{RBM}_{n,m}) \leq n - \lfloor \log(m + 1) \rfloor - \frac{(m + 1)}{2^{\lfloor \log(m+1) \rfloor}} \approx (n - 1) - \log(m + 1)\ .$$

Conversely, given an error tolerance $0 \leq \epsilon \leq 1$, the choice $m \geq 2^{(n-1)(1-\epsilon)+0.1} - 1$ ensures a sufficiently rich RBM model that satisfies $D_{\mathrm{RBM}_{n,m}} \leq \epsilon D_{\mathrm{RBM}_{n,0}}$.

For $m = 2^{n-1} - 1$ the error vanishes, corresponding to the fact that an RBM with that many hidden units is a universal approximator (see Theorem 3.1.1).

Proof of Theorem 4.2.2. From Theorem 4.2.1 we know that $\mathrm{RBM}_{n,m}$ contains the union $\mathcal{M}$ of all mixtures of independent models corresponding to all partitions with up to $m + 1$ cubical blocks. Hence $D_{\mathrm{RBM}_{n,m}} \leq D_{\mathcal{M}}$. Let $k = n - \lfloor \log(m + 1) \rfloor$ and $l = 2m + 2 - 2^{n-k+1} \geq 0$; then $l2^{k-1} + (m + 1 - l)2^k = 2^n$. Lemma 4.1.6 with $n_1 = \cdots = n_l = k - 1$ and $n_{l+1} = \cdots = n_{m+1} = k$ implies

$$D_{\mathcal{M}} \leq \frac{l(k - 2)}{2^{n-k+1}} + \frac{(m + 1 - l)(k - 1)}{2^{n-k}} = k - \frac{m + 1}{2^{n-k}}\ .$$

This completes the proof of the main statement.

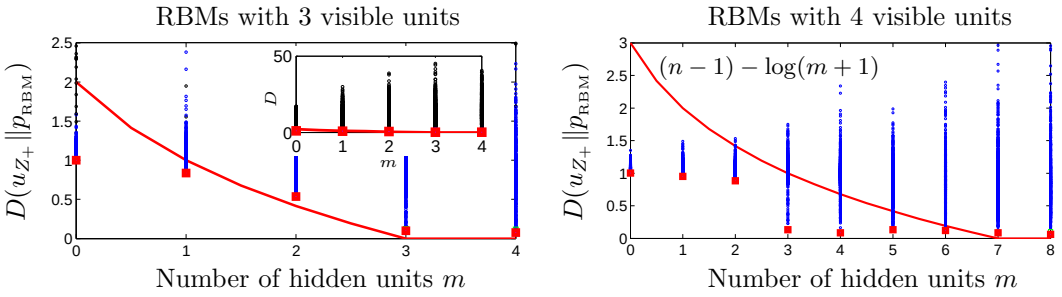

Figure 4.3: *This figure demonstrates Theorem 4.2.2 for $n = 3$ and $n = 4$ visible units. The red curves indicate the bounds from the theorem. We fixed u_{Z_+} as target distribution, the uniform distribution on binary length n vectors with an even number of ones. The inset of the left figure shows the resulting KL-divergence $D(u_{Z_+}\|p^{rand}_{\mathrm{RBM}})$ for a random initialization of the parameters (for $n = 4$ the resulting KL-divergence was larger). After training the RBMs the result (blue circles) is often not better than the uniform distribution, for which $D(u_{Z_+}\|u) = 1$. For each m, the best set of parameters was used to initialize a further CD training with a smaller learning rate (green squares, mostly covered) followed by a short maximum likelihood gradient ascent (red filled squares).*

To see that the choice $m \geq 2^{(n-1)(1-\epsilon)+0.1} - 1$ ensures the given approximation error: The maximal approximation of $\mathrm{RBM}_{n,0}$ is $(n-1)$. Furthermore, elementary analysis shows

$$n - \lfloor \log(m+1) \rfloor - \frac{(m+1)}{2^{\lfloor \log(m+1) \rfloor}} \leq (n-1) - \log(m+1) + c\,,$$

with $c = (-\log(\ln(2)) - (\frac{1}{\ln(2)} - 1)) < \frac{1}{10}$. □

In Figure 4.3 we use computer experiments to illustrate Theorem 4.2.2. For $n = 3$ and $n = 4$ we fixed $u_{Z_{+,n}}$ as target distribution, the uniform distribution on binary length n vectors with an even number of ones. The distribution u_{Z_+} is not necessarily the KL-maximizer from $\mathrm{RBM}_{n,m}$, but it is difficult to represent. Qualitatively, samples from u_{Z_+} look like uniformly distributed, and representing u_{Z_+} requires the maximal number of product mixture components (see Chapter 1 and Appendix 3.B). For both values of n and each $m = 0, \ldots, 2^n/2$ we initialized 500 resp. 1000 RBMs at parameter values chosen uniformly at random in the range $[-10, 10]$. Randomly chosen distributions in $\mathrm{RBM}_{n,m}$ are likely to be very far from the target distribution. We trained these randomly initialized RBMs using contrastive divergence with 500 training epochs, learning rate 1 and a list of even parity vectors as training data. The resulting KL-divergence respects the bound given in Theorem 4.2.2.

4.3 Deep Belief Networks

There are three reasonable approaches to find explicit submodels of $\mathrm{DBN}(n_0^l)$: First, to study the models arising form *probability sharing* on RBMs, as in the constructions of universal approximators provided in Chapter 3, but with a restricted number of sharing steps (i.e., for a specified number of hidden layers and widths). Second, to study the set $\mathrm{DBN}(n_0^l)$ as a mixture of conditional probability distributions with mixing distributions from $\mathrm{DBN}(n_1^l)$. Third, to study the set of joint probability distributions $\mathcal{D}(n_0^l)$ and especially the elements $P(v,h) = p(v)\delta_{h_v}(h)$. In

this section we explore the three approaches for DBNs with layers of equal width. Our approach can be extended to treat DBNs with layers of different sizes. For reasons of simplicity we omit this generalization.

Probability Sharing in Layers of Equal Width

We consider the set of all unions of k *edges* of the n-dimensional unit cube:

$$\mathbb{E}_{k,n} := \{\cup_{j\in[k]}\{x^{(1)}, \hat{x}^{(1)}\} \colon x^{(j)}, \hat{x}^{(j)} \in \{0,1\}^n \text{ and } d_H(x^{(j)}, \hat{x}^{(j)}) = 1\} \,. \tag{4.10}$$

By Theorem 3.1.1, each element of $\mathbb{E}_{k+1,n}$ is an S-set of $\mathrm{RBM}_{n,k}$, i.e., the RBM with n visible and k hidden units can approximate any distribution with support on one of these sets arbitrarily well.

An n-bit Gray code of cardinality l is a collection of l binary vectors of length n such that subsequent elements have Hamming distance one to each other. This can be understood as a path of length l on the graph of the n-dimensional cube. A Gray code is conveniently written as a $(l \times n)$-binary matrix. The *transition sequence* of a Gray code $G = \begin{pmatrix} G_1 \\ \vdots \\ G_l \end{pmatrix} \in \{0,1\}^{l\times n}$ is a list of numbers $T_k \in [n]$ for $k \in [l-1]$ indicating the position where the vectors G_k and G_{k+1} differ. The *vertex visited by the path at time* t is the binary vector G_t. The bit changed a time t is the number T_t.

The following collection of sets consists of unions of paths on the vertices of the n-cube, with *(i)* the starting points of the paths are contained in a set of $\mathbb{E}_{k,n}$, and *(ii)* at some time t any two paths change different bits, unless they are visiting neighboring vertices.

Definition 4.3.1. Let l be a natural number and let $G^{(i)}$ be an n-bit Gray code of length $l^{(i)} \leq l$ with first element $G_1^{(i)} = s^{(i)} \in \{0,1\}^n$ and transition sequence $T^{(i)} \in [n]^{l^{(i)}-1}$ for all $i \in [k]$.

$$\mathbb{S}_k^l := \{G^{(1)} \cup \cdots \cup G^{(r)}\}, \quad \text{where}$$
$$\begin{aligned} &\bullet\ s^{(i)} \neq s^{(j)}\ \forall i \in [r], \text{ and } \cup_{i\in[r]}\, s^{(j)} \subseteq E \in \mathbb{E}_{k,n}, \\ &\bullet\ G_1^{(i)} = s^{(i)} \text{ for all } i \in [r] \\ &\bullet\ T_t^{(i)} \neq T_t^{(j)} \text{ unless } d_H(G_t^{(i)}, G_t^{(j)}) = 1 \end{aligned} \tag{4.11}$$

This definition is motivated by the probability sharing scheme used in Section 3 based on [73]. For any $l \geq 1$, the family $\mathbb{S}_{(n+1)}^l$ contains any union of $(n+1)$ pairs of vectors with Hamming distance one.

Lemma 4.3.2. (*S*-sets of DBNs). *Let* $l \in \mathbb{N}$ *and* $n_1 = n_2 = \cdots = n_l = n$. $\mathrm{DBN}(n_0^l)$ *contains any* $p \in \mathcal{P}_n$ *with* $\mathrm{supp}(p) \in \mathbb{S}_{(n+1)}^l$.

Proof. The result is an adaptation of Theorem 3.2.1 (see pg. 79). The claim follows using Lemma 3.2.5, where we just need to replace "$p^* \in \overline{\mathcal{P}}$" by "$p^* \in \cup_{i,k}\{S_{i,k}\}$" and "$\frac{2^n}{a}$" by "$l$". The requirements of that lemma can be checked using Theorem 3.2.2, Theorem 4.2.1, and the definition of $\mathbb{S}_k^l$ given in eq. (4.11). □

We call the set of vertices incident to a K-dimensional face of the n-dimensional unit cube a *K-face of the n-cube*.

Lemma 4.3.3. *For the collection of sets $\mathbb{S}^l_{(n+1)} \subseteq 2^{\{0,1\}^n}$ we have:*

(i) If $n = R + 2^k + k + 1$ and $l \geq 2^{2^k}$ for some $k \in \mathbb{N}$, then any $(n-R)$-face of the n-cube is in $\mathbb{S}^l_{(n+1)}$.

(ii) If $n \geq N \cdot K$, $K = 2^k + k + 1$, and $l \geq 2^{2^k}$ for some $k \in \mathbb{N}$, then $\mathbb{S}^l_{(n+1)}$ contains any union of N K-cylinders $[y_{[n]\setminus\lambda_i}]$ with disjoint sets $\lambda_i \subseteq [n]$, $i = 1, \ldots, N$.

(iii) If $k < n$ and $l \geq 2^k$ for some $k \in \mathbb{N}$, then $\mathbb{S}^l_{(n+1)}$ contains any $(k + \log 2k)$-face of the n-cube.

Proof. (i) Use [87, Lemma 1]. *(ii)* Use the first part. *(iii)* If $k \leq n$ and $l \geq 2^k$ for some $k \in \mathbb{N}$, then $\mathbb{S}^l_{(n+1)}$ contains the union of any k cylinder sets $[y_{[n]\setminus\lambda}]$, $|\lambda| = k$. Use a k-bit Gray code G on the bits λ (corresponding to any face with fixed values on $[n] \setminus \lambda$). The entries $[n] \setminus \lambda$ are arbitrary. The k codes satisfy $T^i_k \neq T^j_k$ if we shift the columns of G cyclically. If $k < n$ and $l \geq 2^k$ for some $k \in \mathbb{N}$, then $\mathbb{S}^l_{(n+1)}$ contains the union of $2k$ k-dimensional cylinder sets $[y_{[n]\setminus\lambda}]$, $[\tilde{y}_{[n]\setminus\lambda}]$, $|\lambda| = k$, $H(y_{[n]\setminus\lambda}, \tilde{y}_{[n]\setminus\lambda}) = 1$. We use the same Gray codes for neighboring $y_{[n]\setminus\lambda}$ and $\tilde{y}_{[n]\setminus\lambda}$. Clearly, these codes change the same bit iff they are visiting neighboring points. Any $(k + \log 2k)$-dimensional cube can be decomposed as a union of that form. □

Lemma 4.3.2 and Lemma 4.3.3 describe S-sets of DBNs. Notice that $\mathbb{S}^l_k$ is not necessarily inclusion complete, and yet if $\mathcal{Y} \in \mathbb{S}^l_k$, then all subsets of $\mathcal{Y}$ are S-sets of the DBN. A particular instance of Lemma 4.3.3 item *(i)* is the following: If $n = 2^k + k + 1$ and $l \geq 2^{2^k}$ for some $k \in \mathbb{N}$, then $\{0,1\}^n \in \mathbb{S}^l_{(n+1)}$, which corresponds to [87, Lemma 1] and recovers precisely the result abut universal approximators [87, Theorem 2].

In order to compute meaningful approximation error bounds for DBNs we need to account for full support distributions contained in this model, since $D(q\|p) = \infty$ whenever $\operatorname{supp}(q) \supsetneq \operatorname{supp}(p)$. This is done in the following subsection:

Hierarchical Mixtures

Proposition 4.3.4. *Consider any $\{n_k \in \mathbb{N}\}_{k=-1}^l$. $\mathrm{DBN}(n^l_{-1})$ is the closure of the set of distributions of the following form:*

$$p(v) = \sum_{h \in \{0,1\}^{n_0}} \frac{\exp((hW + b)v)}{Z_{hW+b}} q(h) \,, \tag{4.12}$$

where $q \in \mathrm{DBN}(n^l_0) \subseteq \overline{\mathcal{P}_{n_0}}$, $W \in \mathbb{R}^{n_0 \times n_{-1}}$ and $b \in \mathbb{R}^{n_{-1}}$.

Proof. From eq. (3.4) we have that $P(h^l_{-1}) = P(h^l_0)P(h^{-1}|h^0)$. Furthermore, $\sum_{h^l_1} P(h^l_{-1}) =$

$P(h^{-1}|h^0)\sum_{h_1^l} P(h_0^l) = P(h^{-1}|h^0)q$ and

$$P(h^{-1}|h^0) = \prod_{j=1}^{n-1} P((h^{-1})_j|h^0) = \prod_{j=1}^{n-1} \frac{\exp((b_j + h^0 W_{j,:})(h^{-1})_j)}{1+\exp(b_j + h^0 W_{j,:})}$$
$$= \frac{\exp(\sum_j (b_j + h^0 W_{j,:})(h^{-1})_j)}{\prod_j (1+\exp(b_j + h^0 W_{j,:}))} = \frac{\exp((b + h^0 W)h^{-1})}{\sum_{h^{-1}} \exp((b + h^0 W)h^{-1})} ,$$

where $W_{j,:}$ denotes the j-th row of W. □

Consider any $n, m \in \mathbb{N}$, as well as $W \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^n$, and the set $\{p_h\}_h$, defined in the following way:

$$p_h \in \mathcal{P}_n : \quad p_h(v) = \frac{\exp((h^\top W + b^\top)v)}{Z_{(hW+b)}} \quad \forall h \in \{0,1\}^m . \tag{4.13}$$

Every such p_h is a product distribution of n binary variables and hence, Proposition 4.3.4 describes a mixture of products. The difficulty involved here is that the p_h share the parameters W and b and hence the set $\{p_h\}_{h\in\{0,1\}^m}$ is not an arbitrary (2^m)-tuple of product distributions. More precisely, only the (2^m)-tuples of product distributions with natural parameters given by the Minkowski sums of the origin and m points in $\mathbb{R}^n$, the rows of W, shifted by the vector b are allowed. By Lemma 4.3.2, the mixing distribution q in eq. (4.12) can be chosen arbitrarily, contingent to having a limited support.

Let e_i be the binary vector for which only the i-th entry is equal to one.

Lemma 4.3.5. *Consider the set $\{p_h\}$ from eq.* (4.12)*. Any of the following items is satisfied to an arbitrary accuracy for a corresponding choice of W and b:*

(i) *$p_{h=e_i}$, $i \in [m]$, are any m product distributions (on arbitrary faces of the n-cube).*

(ii) *Let C be any K-dimensional face of the m-cube, for any $K \leq m$. $\{p_h\}_{h\in C}$ contains all uniform distributions supported in the flats (intersection semilattice) of K arbitrary faces of the n-cube.*

(iii) *Let C be any K-dimensional face of the m-cube for some $K \leq m, n$. Consider any $\lambda \subseteq [n]$ with $|\lambda| = K$. The sets $\{x\colon x_\lambda = h\}_{h\in C}$ have cardinality 2^{n-K} and build a partition of $\{0,1\}^n$. The element p_h is the uniform distribution on $\{x\colon x_\lambda = h\}$ for every $h \in C$.*

(iv) *Let $m = n$ and consider any m disjoint edges of the m-cube $\{f^i, g^i\}_{i\in[m]}$. p_{f^i} is arbitrary on $\{f^i, f^i + e_i \mod 2\}$, p_{g^i} is arbitrary on $\{g^i, g^i + e_i \mod 2\}$, and p_h is the point measure on $\{v = h\}$ for all $h \notin \{f^i, g^i\}$.*

Proof. *(i)* Consider a face $F = \{x\colon x_\lambda = y_\lambda\}$, where $\lambda \subset [n]$. Choosing $W_{i,:} = \lim_{\alpha\to\infty}(\alpha(y_\lambda - \frac{1}{2}\mathbb{1}), \beta_{[n]\setminus\lambda})$ results in a $p_{h=e_i}$ which has support F and $p_{h=e_i}(y_\lambda, x_{[n]\setminus\lambda}) \propto \exp(\beta_{[n]\setminus\lambda} \cdot x_{[n]\setminus\lambda})$ for $(y_\lambda, x_{[n]\setminus\lambda}) \in F$.

(ii) To simplify the notation we identify a vector h with its support set (e.g., e_i is identified with $\{i\}$). Choose parameters such that $p_{\{i\}}$ is a uniform distribution on the face F_i of the n-cube for all $i \in [m]$. p_λ is uniformly distributed with support $\operatorname{argmax}(\sum_{i\in\lambda} e_{F_i})$. Hence, if $F_\lambda := \cap_{i\in\lambda} F_i \neq \emptyset$, then $\operatorname{supp}(p_\lambda) = F_\lambda$.

(iii) Let $W_{:,\lambda}$ denote the entries of W in the columns $j \in \lambda$. The statement follows from choosing $b = -\alpha\mathbb{1}$, $W_{:,\lambda} = \alpha E_\lambda$ (where E_λ the unit matrix), and $W_{:,[n]\setminus\lambda} = 0$.

(iv) Consider any $l \in [n]$. Consider a pair of vectors $\{f, g\}$ which is an edge of $\{0,1\}^m$. Let $r \in [m]$ be the entry where they differ. Let $s \in [m]$ be arbitrary. Denote by $\hat{f}$ the vector $\hat{f}_i = f_i \, \forall i \neq r, s$ and $\hat{f}_r = 0$, $\hat{f}_s = 0$. Choosing

$$\begin{aligned} W_{:,l} &= \omega(2\hat{f} - \hat{\mathbb{1}} + (1 - 2f_s)me_s + (p - q)e_r) \\ b_l &= -\omega(|\mathrm{supp}(f)| - 1 + f_s m) + q \end{aligned}$$

yields in the limit $\omega \to \infty$ that $P(v_l = h_s | h \neq f, g) = 1$, $P(v_l = 1 | h = f) = p$, and $P(v_l = 1 | h = g) = q$, i.e.,

$$\begin{aligned} P(v_l | h \neq f, g) &= \delta_{h_s}(v_l) \\ P(v_l | h = f) &= p(v_l) \\ P(v_l | h = g) &= q(v_l) \, . \end{aligned}$$

Consider the case $m = n$. Let $\{f^i, g^i\}_{i=1}^m$ be m disjoint edges of $\{0,1\}^m$. Let $s^i = i \; \forall i \in [m]$. Consider any $l \in [n]$. From the above discussion we get

$$P(v | h = f^l) = \prod_{i=1}^n P(v_i | f^l) = \prod_{i \neq l} \delta_{f^l_{s^i}}(v_i) \cdot p^l(v_l) \, , \tag{4.14}$$

which is an arbitrary distribution with support on the edge given by fixing $v_i = f^l_i \, \forall i \neq l$. For $h \notin \cup_{i=1}^m \{f^i, g^i\}$ and $s^i = i \, \forall i$ we get

$$P(v | h \neq f^l, g^l \, \forall l) = \prod_{i=1}^n P(v_i | h) = \prod_i \delta_{h_{s^i}}(v_i) = \delta_h(v) \, , \tag{4.15}$$

which is the point measure on $\{v = h\}$. □

Main Results

Theorem 4.3.6. *Consider a DBN containing l hidden layers of the same width as the visible layer, n. Let k be the largest natural number for which $l - 1 \geq 2^{2^k}$ and let $K = 2^k + k + 1 \leq n$. The DBN model contains:*

- *Any $p \in \overline{\mathcal{P}_n}$ with support contained in an element of $\mathbb{S}^l_{(n+1)}$.*
- *Any partition model with blocks of the form $\{[y_\lambda]\}_{y_\lambda \in \{0,1\}^K}$, $\lambda \subseteq [n], |\lambda| = K$.*

Proof. The result summarizes statements from Lemma 4.3.2 and Lemma 4.3.5. □

By this result, if $K \geq n$, the DBN is a universal approximator, which is consistent with Theorem 3.2.1. Since the DBN contains $\mathcal{P}(\mathcal{Y})$ for all $\mathcal{Y} \in \mathbb{S}^l_{(n+1)}$, the cardinality of any such $\mathcal{Y}$ (minus one) is a lower bound on the dimension of the DBN model.

Theorem 4.3.7. *The DBN from Theorem 4.3.6 satisfies*

$$\max_{p \in \mathcal{P}_n} D(p \| \, \mathrm{DBN}) \leq n - K \, ,$$

where $K = 2^k + k + 1 = \log(2l \log(l))$.

Proof. The claim follows from Corollary 4.1.1 and the second item of Theorem 4.3.6. □

The DBN Joint Model

In this short passage we discuss a third approach to the expressive power of DBNs (the other two being the previously discussed probability sharing and hierarchical mixtures). Consider any natural numbers $\{n_0, \dots, n_l\}$ and the corresponding architecture of $\mathcal{D}(n_0^l)$. Let $\mathcal{E}^1 \subseteq \mathcal{P}_N$ denote the set of product distributions of N units. Let $\mathcal{E}^*$ be the hierarchical model with interaction family Δ consisting of $\{i\}$ for all $i \in N$, the pairs $\{i, j\}$, where i and j belong to subsequent layers of the DBN, and additionally, all subsets of units from the same hidden layer $k, k \in \{1, \dots, l-1\}$. We have the following:

Proposition 4.3.8. $\mathcal{E}^1 \subset \mathcal{D}(n_0^l) \subseteq \mathcal{E}^*$.

Proof. The first inclusion is clear. For the second inclusion: $P \in \mathcal{D}$ is of the form

$$P(h_0^l) = \exp\big(h^l W^l h^{l-1} + b^l h^l + b^{l-1} h^{l-1} - \log Z_l \\ + \sum_{k=0}^{l-2} (h^{k+1} W^{k+1} + b^k) h^k - \log Z_{k+1}(h^{k+1})\big) \ ,$$

where $Z_l = \sum_{h^l, h^{l-1}} \exp(h^l W^l h^{l-1} + b^l h^l + b^{l-1} h^{l-1})$ is a constant, and $Z_{k+1}(h^{k+1}) = \sum_{h^k} \exp((h^{k+1} W^{k+1} + b^k) h^k)$ is a function of h^{k+1}. From the definition of $\mathcal{E}^*$, all these functions must be contained in the span of the sufficient statistics of $\mathcal{E}^*$. □

The functions $\log Z_{k+1}$ do not necessarily belong to any specific subspace of the span of the sufficient statistics of $\mathcal{E}^*$. This suggests that any exponential family containing $\mathcal{D}$ also contains $\mathcal{E}^*$.

4.A A Comparison of Restricted Boltzmann Machines and Mixture Models

RBMs generate mixtures of product distributions, according to the relation

$$p = \sum_{h \in \{0,1\}^m} p(v|h) p(h) \ , \text{ where } p(v|h) \in \mathcal{E}_n^1 \ , \tag{4.16}$$

and $p(h)$ is the marginal distribution on the hidden states $\{0, 1\}^m$ of the RBM. This implies $\mathrm{RBM}_{n,m} \subseteq \mathrm{Mixt}^{2^m}(\overline{\mathcal{E}_n^1})$. The mixtures produced by the RBM have restricted weights $p(h)$ and restricted mixture components $p(\cdot|h)$. In general $\mathrm{RBM}_{n,m} \neq \mathrm{Mixt}^{2^m}(\overline{\mathcal{E}_n^1})$, as can be seen from dimension arguments (the mixture model has dimension $2^m n + 2^m - 1$ for all m, see [26]). In the particular case $m = 1$ we have (by Theorem 4.2.1):

$$\mathrm{RBM}_{n,1} = \mathrm{Mixt}^2(\overline{\mathcal{E}_n^1}) \quad \forall n \in \mathbb{N} \ . \tag{4.17}$$

The statement of eq. (4.17) also appeared in [31, Proposition 3.1]. More generally [31] shows that $\mathrm{RBM}_{n,m} = (\mathrm{RBM}_{n,1})^{[m]}$. Given some set $\mathcal{M} \subset \mathcal{P}$, $\mathcal{M}^{[m]}$ is the m-th *Hadamard product* of $\mathcal{M}$, i.e., the renormalized entry-wise product of m probability distributions in $\mathcal{M}$. In this sense an RBM model is a *product of mixtures*. Hinton [55] places RBMs in a class of models called *product of experts*, where the experts belong to $\mathrm{Mixt}^2(\mathcal{E}_n^1)$.

If we fix the number of parameters, are RBM models better than mixtures of products at approximating interesting probability distributions? Hinton [55] explains advantages of products of experts compared to *mixtures of experts* (directed acyclic graphical models in that context). It is well known that RBMs create a *multi-clustering* [17, Section 5.3], where each hidden unit linearly divides the input space, and the resulting partition corresponds to input regions where $p(h|x)$ is maximized by different hidden states. However, not every hidden state corresponds to a non-empty region of the input space [17]. What kind of partitions are possible, and how do they relate to the modes that an RBM can realize on the input states?

Question 4.A.1. Given some $n, m \in \mathbb{N}$, what is the smallest k for which the k-mixture of product distributions contains the RBM model, $\min\{k\colon \mathrm{Mixt}^k(\mathcal{E}_n^1) \supseteq \mathrm{RBM}_{n,m}\}$?

In the following we explore the kinds of modes that can be realized by probability distributions in an RBM model. Then we use that information to assess Question 4.A.1 and show that in essentially all cases of practical interest, an exponentially larger mixture model, requiring an exponentially larger number of parameters, is required to represent the distributions that can be represented by the RBM (Proposition 4.A.13).

We start with some general observations about the joint probability distributions on the visible and hidden states of an RBM:

Sufficient Statistics of the RBM Joint Model

The joint probability distributions on the states of hidden and visible units of $\mathrm{RBM}_{n,m}$ have the following form:

$$p_{W,C,B}(v,h) = \frac{1}{Z_{W,C,B}} \exp\left(h^\top W v + C^\top h + B^\top v\right) \quad \forall (v,h) \in \{0,1\}^{n+m} .$$

Denote by $K_{n,m}$ the full bipartite graph with $n+m$ vertices. This defines the hypergraph of interactions of $\mathrm{RBM}_{n,m}$. The RBM joint model is $\mathcal{E}_{K_{n,m}} \subseteq \mathcal{P}_{n+m}$, the hierarchical model with interaction sets $K_{n,m}$.

Denote by $\underline{h}$ the matrix with the states of the hidden units as columns $\underline{h} = (h^1|\cdots|h^{2^m}) \in \{0,1\}^{m\times 2^m}$, and $\underline{v} \in \{0,1\}^{n\times 2^n}$ the matrix with the states of the visible units as columns. Denote by $\mathbb{1}_{a\times b}$ the $a \times b$ matrix with all entries equal to one. The sufficient statistics of $\mathcal{E}_{K_{n,m}}$ is

$$T^{n,m} = \begin{pmatrix} \mathbb{1}_{1\times 2^{n+m}} \\ \mathbb{1}_{1\times 2^m} \otimes \underline{v} \\ \underline{h} \otimes \mathbb{1}_{1\times 2^n} \\ \underline{h} \otimes \underline{v} \end{pmatrix} , \tag{4.18}$$

where the columns are indexed by $(h^1, v^1), (h^1 v^2), \ldots, (h^1, v^{2^n}), \ldots, (h^{2^m}, v^{2^n})$. The block $\begin{pmatrix} \mathbb{1}_{1\times 2^m} \otimes \underline{v} \\ \underline{h} \otimes \mathbb{1}_{1\times 2^n} \end{pmatrix}$ is the sufficient statistics of the independence model on $(n+m)$ binary variables. The last block represents the interactions between pairs of visible and hidden units. The convex support $\mathrm{cs}(\mathcal{E}_{K_{n,m}})$ is a $(nm+n+m)$-dimensional polytope (see Chapter 1). The face lattice of this polytope corresponds to the support sets of probability distributions contained in $\overline{\mathcal{E}_{K_{n,m}}}$ and the simplex faces correspond to faces of $\mathcal{P}$ which are contained in $\overline{\mathcal{E}_{K_{n,m}}}$. The RBM joint model contains the independence model and is contained in the pairwise interaction model.

It appears natural to use this information to assess the expressive power of RBMs. This approach presents difficulties, mainly because the face lattice of the convex support is itself a complicated object and not sufficiently well understood.

Modes of Distributions in the RBM

Corollary 1.3.3 and eq. (4.16) imply the following:

Proposition 4.A.2. *If $\mathrm{RBM}_{n,m}$ can represent some $p \in \mathcal{P}(C)$, where C is an n-bit code of minimum distance two, then $\mathcal{E}_{K_{n,m}}$ has a facial set $\{(x^i, y^{i,j})\}_{j,i} \subseteq \{0,1\}^{n+m}$ satisfying $\cup_i x^i = C$ and $y^{i,j} \neq y^{i',j'}$ for all j, j' whenever $x^i \neq x^{i'}$. In particular:*

(i) The model $\mathrm{RBM}_{n,m}$ doesn't contain any probability distribution supported by an n-bit code C of minimum distance two and cardinality $|C| > 2^m$. If $m \leq n-2$, then $\mathrm{RBM}_{n,m}$ doesn't contain any distribution supported by $Z_{\pm,n}$.

(ii) If $m = n-1$ and $\mathrm{RBM}_{n,m} \supset \mathcal{P}(Z_{\pm,n})$, then $\mathcal{E}_{K_{n,m}}$ has an S-set $\mathcal{Y} \subseteq \{0,1\}^{n+m}$ with $\mathcal{Y}_{[n]} := \{y \in \{0,1\}^n \colon \exists z \in \{0,1\}^m \text{ with } (y,z) \in \mathcal{Y}\} = Z_{\pm,n}$.

Consider the particular case $m = n-1$. If $\mathrm{RBM}_{n,n-1}$ is a universal approximator, then $n - 1 \geq \frac{2^n - n - 1}{n+1}$ (parameter counting, see Corollary 3.A.2). Now, if $\mathrm{RBM}_{n,n-1}$ is a universal approximator, then it can represent $\mathcal{P}(Z_{\pm,n})$ and by Proposition 4.A.2 there is some S-set $\mathcal{Y} \subset \{0,1\}^{n+(n-1)}$ of $\mathcal{E}_{K_{n,n-1}}$ which restricts to $Z_{+,n}$ on the visible entries. In this case $\mathcal{Y}$ is a facial set and the matrix $T_{\mathcal{Y}}^{n,n-1}$ (which consists of the columns $\mathcal{Y}$ of the sufficient statistics $T^{n,n-1}$) has full rank 2^{n-1}. We computed the rank of $T_{\mathcal{Y}}^{n,n-1}$ for a set $\mathcal{Y} \in \{0,1\}^{n+(n-1)}$ which restricts to $Z_{+,n}$ in the visible variables and to $\{0,1\}^m$ in the hidden variables, for several values of n:

n	2	3	4	5	6	7	8	9	10
$\mathrm{rk}(T_{\mathcal{Y}}^{n,n-1})$	2	4	8	16	27	56	72	90	110
2^{n-1}	2	4	8	16	32	64	128	256	512
$n-1 \geq \frac{2^n-n-1}{n+1}$	yes	yes	yes	no	no	no	no	no	no

We see that $\mathrm{RBM}_{n,n-1}$ is not a universal approximator for $n \geq 5$, and that it does not contain $\mathcal{P}(Z_{\pm,n})$ for $n \geq 6$. For $n \leq 3$, the number of parameters and the rank of the matrix do not contradict $\mathcal{P}(Z_{\pm,n}) \subset \mathrm{RBM}_{n,n-1}$. However, we still have to check that $\mathcal{Y}$ is a facial set. We will return to this later in this section.

The following definition captures properties of probability distribution that we will use in Proposition 4.A.7.

Definition 4.A.3. For any $\mathcal{C} \subset \{0,1\}^n$ we define $\mathcal{I}_{\mathcal{C}} \subset \overline{\mathcal{P}_n}$ as the set of probability distributions for which any representation as mixture of product distributions contains $|\mathcal{C}|$ components with unique maximizers at the different elements of $\mathcal{C}$:

$$\mathcal{I}_{\mathcal{C}} := \Big\{p \in \overline{\mathcal{P}_n} \colon p = \sum_i \alpha_i p^i,\ p^i \in \overline{\mathcal{E}_n^1} \Rightarrow \text{ for each } x \in \mathcal{C} \text{ there is some } p^i \text{ with } \operatorname{argmax}(p^i) = x\Big\}\ .$$

Example 4.A.4. If $\mathcal{C} \subset \{0,1\}^n$ is a binary code of minimum Hamming distance at least two, then by Lemma 1.B.8 the set $\mathcal{H}_{\mathcal{C}}$ of distributions with strong modes $\mathcal{C}$ is a subset of $\mathcal{I}_{\mathcal{C}}$. In particular $\mathcal{H}_{n,2^{n-1}}$, the set of distributions with 2^{n-1} strong modes on $Z_{\pm,n}$, is a subset of $\mathcal{I}_{Z_{\pm,n}}$. We write $\mathcal{I}_n^{\pm}$ for $\mathcal{I}_{Z_{\pm,n}}$.

Example 4.A.5. In the case $n = 2$ we have that the set $\mathcal{I}_2^{\pm}$ is equal to $\mathcal{G}_2^{\pm}$; the set of distributions $p = (p(x))_{x\in\mathcal{X}} \in \overline{\mathcal{P}_2}$ with two modes, i.e., the set of distributions for which $p(z) > p(y)$ for all $z \in Z_{\pm}$ for all y with $d_H(x, y) = 1$. In this case $\mathcal{G}_2^{\pm} = \bigcap_{x\in Z_{\pm}} V_x \cap \overline{\mathcal{P}}$, where $V_x \subset \mathbb{R}^{\mathcal{X}}$ is the cell of δ_x in the *Voronoi diagram* of $\mathbb{R}^{\mathcal{X}}$ with centers $\{\delta_v\}_{v\in\{x\}\cup Z_{\mp}}$. $\mathcal{G}_2^{\pm}$ is the convex hull of $\{\delta_x\}_{x\in Z_{\pm}} \cup C$, where C is the set of centroids of the faces of $\overline{\mathcal{P}_2}$ which contain $\{\delta_x\}_{x\in Z_{\pm}}$. See also Figure 1.5.

If $\mathcal{C}$ is a code of minimum distance two, then the mixture model of product distributions $\mathrm{Mixt}^k(\mathcal{E}_n^1)$ intersects $\mathcal{I}_{\mathcal{C}}$ if and only if $k \geq |\mathcal{C}|$, see Lemma 1.B.8. In the following we give a condition for RBM models to intersect $\mathcal{I}_{\mathcal{C}}$.

Definition 4.A.6. Given any $m, n \in \mathbb{N}$ and vectors $\{w_i\}_{i\in[m]} \subset \mathbb{R}^n$ and $b \in \mathbb{R}^n$, the set $\mathcal{Z} := \mathrm{conv}(\{b + \sum_{i\in I} w_i\}_{I\subseteq[m]})$ is an *m-generated zonotope*. The set $\{b + \sum_{i\in I} w_i\}_{I\subseteq[m]}$ is called the set of *points* of $\mathcal{Z}$.

Zonotopes have many interesting properties; in particular, they can be identified with hyperplane arrangements and oriented matroids, see [21, 122]. An orthant of $\mathbb{R}^n$ is the set of all vectors in $(\mathbb{R} \setminus \{0\})^n$ which have the same sign in each entry. Each orthant is labeled by its sign vector. We say that an orthant has even (odd) parity if its sign vector has an even (odd) number of $+$. $\mathbb{R}^n$ has 2^n orthants, 2^{n-1} even and 2^{n-1} odd. A binary vector x is identified with the sign vector $\mathrm{sgn}(x - \frac{1}{2})$.

Proposition 4.A.7. *Let $\mathcal{C}$ be a binary code of minimum distance two. If the model $\mathrm{RBM}_{n,m}$ contains some $p \in \mathcal{I}_{\mathcal{C}}$, then there is a zonotope $\mathcal{Z}$ in $\mathbb{R}^n$ with m generators such that the set of points of $\mathcal{Z}$ intersects every orthant of $\mathbb{R}^n$ of sign $\mathcal{C}$. If $|\mathcal{C}| = 2^m$ and there exists a zonotope with points of equal l_1 norm, which intersect the orthants $\mathcal{C}$ of $\mathbb{R}^n$, then $\mathrm{RBM}_{n,m}$ contains $u_{\mathcal{C}}$ (and intersects $\mathcal{I}_{\mathcal{C}}$).*

Proof. If some $p \in \mathcal{I}_C$ is contained in $\mathrm{RBM}_{n,m}$, then for each $x \in \mathcal{C}$ there is an $h \in \{0, 1\}^m$ such that $p(v|h) \propto \exp(hWv + Bv + Ch)$ is uniquely maximized at $v = x$. Hence

$$hWx + Bx + Ch > hWv + Bv + Ch \quad \forall v \neq x\ , \tag{4.19}$$

and equivalently, $\mathrm{sgn}(hW + B) = \mathrm{sgn}(x - \frac{1}{2})$. Therefore, the existence of a zonotope as described in the claim of this proposition is equivalent to the satisfiability of the inequalities (4.19). On the other hand, if the inequalities (4.19) are satisfied for a set of parameters W, B, C and all vectors $hW + B$ have the same one-norm, then the parameters $\alpha W, \alpha B$, $C = -W(1, \ldots, 1)^\top$, and $\alpha \to \infty$ produce $u_{\mathcal{C}}$ as visible distribution of the RBM. □

Remark 4.A.8. The model $\mathrm{RBM}_{n,m}$ is symmetric under relabeling of the variable x_1. If $\mathrm{RBM}_{n,m}$ intersects $\mathcal{I}_n^+$, then it also intersects $\mathcal{I}_n^-$.

The convex hull of a pair $b_1 \in \mathbb{R}^2_+$ and $b_2 \in \mathbb{R}^2_-$ is a one-generated zonotope whose points intersect all even orthants of $\mathbb{R}^2$. We will show that if n is an odd number larger than one, there don't exist $(n - 1)$-generated zonotopes with points intersecting all even orthants of $\mathbb{R}^n$ (Propostion 4.A.12). We will show however that such a zonotope does exist for $n = 4$, and we will use it as "building block" to construct zonotopes in $\mathbb{R}^{4k}$ which intersect many even orthants of $\mathbb{R}^{k4}$.

Definition 4.A.9. A *hyperplane arrangement* $\mathcal{A}$ in $\mathbb{R}^n$ is a finite set of (affine) hyperplanes $\{H_i\}_{i\in[k]}$ in $\mathbb{R}^n$. Choosing an orientation for each hyperplane yields for each vector $x \in \mathbb{R}^n$ a sign vector $\mathrm{sgn}_{\mathcal{A}}(x) \in \{+, -, 0\}^k$, where $(\mathrm{sgn}_{\mathcal{A}}(x))_i$ indicates whether x lies on the positive

side of H_i, on its negative side, or inside. The set of all vectors in $\mathbb{R}^n$ with the same sign vector are called a *cell* of $\mathcal{A}$, see [21].

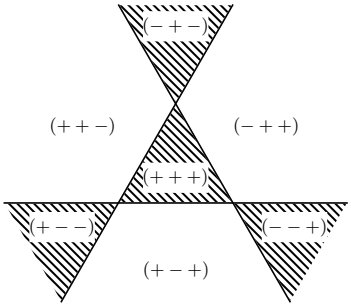

Figure 4.4: *Intersection of a 2-dimensional affine subspace of* $\mathbb{R}^3$ *with normal vector* (111) *and* 7 *orthants of* $\mathbb{R}^3$*; four of odd parity and three of even parity.*

Given some set $\mathcal{C} \subseteq \{\pm 1\}^n$, if there is an m-generated zonotope with points in the orthants $\mathcal{C}$ of $\mathbb{R}^n$, then the vertices of an m-cube are in the $\mathcal{C}$-cells of an arrangement of n hyperplanes in $\mathbb{R}^m$. A subset $\mathcal{C} \subset \{\pm 1\}^m \subset \mathbb{R}^m$ is *linearly separable* iff there exists an affine hyperplane H in $\mathbb{R}^m$ such that $\mathcal{C}$ lies on one side of H and $\{\pm 1\}^m \setminus \mathcal{C}$ lies on the other side of H.

In the following examples we label the vertices of the m-cube by the decimal number that they represent, plus one. A linear separation of the vertices is written as an array of decimal numbers from 1 to 2^m with a bar above the vertices which lie below the separating hyperplane (this notation is common for *covectors* in the context of oriented matroids).

Example 4.A.10. Let $n = 3$ and $m = 2$. If there exists a 2-generated zonotope with vertices intersecting all even orthants of $\mathbb{R}^3$, then there is an arrangement of three hyperplanes $\mathcal{A}$ in $\mathbb{R}^2$ (each hyperplane is just a line) such that the vertices of the 2-cube are in the even cells of $\mathcal{A}$. In this case clearly, each hyperplane separates the vertices of the 2-cube into two sets of cardinality two. There are only two ways to linearly separate the vertices of the 2-cube into sets of cardinality two (up to opposites):

$$12\overline{34} \quad \text{and} \quad 1\overline{2}3\overline{4}\,.$$

Hence there does not exist a 2-generated zonotope with vertices in the four orthants of even (odd) parity of $\mathbb{R}^3$. In combination with Proposition 4.A.7 this shows that $\mathrm{RBM}_{3,2}$ can't represent distributions with four strong modes. The next Proposition 4.A.12 generalizes this example. Later we will study the special case $n = 3$ and $n = 2$ in more detail.

Example 4.A.11. Let $n = 4$ and $m = 3$. There are 104 ways to linearly separate the vertices of the 3-cube, see [89]. A complete list can be found in [21, Section 3.8]. The vertices of the 3-cube are in the $Z_{+,4}$ cells of an arrangement of four hyperplanes which separate the vertices as follows:

$$1234\overline{5678}\,,\ 123\overline{4}56\overline{78}\,,\ \overline{1}234\overline{5}67\overline{8}\,,\ \overline{1}23\overline{4}5\overline{6}7\overline{8}\,.$$

Hence there is a 3-generated zonotope with vertices in the 8 orthants of even (or odd) parity of

$\mathbb{R}^4$. In fact, the following parameters

$$w = \begin{pmatrix} -1 & -1 & -1 & 1 \\ -1 & -1 & 1 & -1 \\ -1 & 1 & -1 & -1 \end{pmatrix} , \quad b = \frac{1}{2}\begin{pmatrix} 3 & 1 & 1 & 1 \end{pmatrix} , \tag{4.20}$$

generate a zonotope in $\mathbb{R}^4$ with the following points:

$$\frac{1}{2}\begin{pmatrix} 3 & 1 & 1 & 1 \\ 1 & -1 & -1 & 3 \\ 1 & -1 & 3 & -1 \\ 1 & 3 & -1 & -1 \\ -1 & -3 & 1 & 1 \\ -1 & 1 & -3 & 1 \\ -1 & 1 & 1 & -3 \\ -3 & -1 & -1 & -1 \end{pmatrix} .$$

Every even orthant of $\mathbb{R}^4$ contains exactly one of these row vectors. This is a contrast to the previous example. The convex hull of the vertices of the 4-dimensional unit cube which have an even number of ones is (combinatorially equivalent to) the 4-dimensional *cross polytope* (the dual of the 4-cube). The n-dimensional cross polytope is defined as $\operatorname{conv}\{(0,\ldots,0,\underset{i}{\pm 1},0,\ldots,0)\}_{i\in[n]}$. The convex hull of the vertices of the d-dimensional unit hypercube which have an even number of ones is called the *parity polytope*. It has dimension d when $d \geq 3$.

Proposition 4.A.12. *If n is an odd natural number larger than one, there is no $(n-1)$-generated zonotope with points intersecting all even (odd) orthants of $\mathbb{R}^n$. In particular, the model $\mathrm{RBM}_{n,n-1}$ can't represent distributions with strong modes on $Z_{\pm,n}$.*

Proof. Let $\mathcal{Z}$ be a candidate zonotope. Since $\mathcal{Z}$ has $n-1$ generators, it has dimension at most $n-1$ and lies in a hyperplane H of $\mathbb{R}^n$. Let η denote the normal vector of that hyperplane. Assume first that H contains the origin. All vectors with sign $\operatorname{sgn}(\eta)$ lie outside of H (where we may assign arbitrary sign on the zero entries of η). This is Stiemke's theorem, see [44]. By linear algebra, the opposite orthant with sign vector $-\operatorname{sgn}(\eta)$ also lies outside H. Since n is odd, one of the two opposite orthants has even parity and the other has odd parity. Hence $\mathcal{Z}$ can't intersect every orthant of even (or odd) parity. Consider now an affine hyperplane H which intersects all even orthants. Assume wlog that the normal vector of H has only negative entries. The intersection $H \cap \mathbb{R}^n_{(-\cdots-)}$ is a (bounded) $(n-1)$-simplex. The orthant $\mathbb{R}^n_{(-\cdots-)}$ is separated by $(n-1)$ hyperplanes from the orthants $\mathbb{R}^n_{s_i}$ with signs $s_i = (+\cdots+\underset{i}{-}+\cdots+)$ for $i \in [n]$. Since n is odd and larger than one, $(n-1)$ is a positive even number. Any collection of n points which intersects $H \cap \mathbb{R}^n_{s_i}$ for all $i \in [n]$ contains $H \cap \mathbb{R}^n_{(-\cdots-)}$ in its convex hull. On the other hand, any $(n-1)$-generated zonotope $\mathcal{Z}$ of dimension $(n-1)$ is combinatorially equivalent to the $(n-1)$-cube. In particular, all points of $\mathcal{Z}$ are vertices. □

See Figure 4.4 for a small example of the objects discussed in Proposition 4.A.12.

The Smallest Mixtures of Products Containing the RBM Model

An analysis of the (strong) modes of probability distributions within an RBM model allows us to derive inclusion relations with mixtures of independence models. We show that, in general,

small RBMs can represent probability distributions with many strong modes, which are only contained in very large mixtures of product distributions.

Starting from a small zonotope with points in all even orthants of $\mathbb{R}^n$ one can construct a larger zonotope with points intersecting many even orthants of a larger space (see Proposition 4.A.7):

For each $i \in [k]$ let n_i, m_i be some natural numbers and $W^{(i)} \in \mathbb{R}^{m_i \times n_i}$, $B^{(i)} \in \mathbb{R}^{n_i}$. If for every i the points of the zonotope generated by $(W^{(i)}, B^{(i)})$ intersect $K_i \in \mathbb{N}$ even (odd) orthants of $\mathbb{R}^{n_i}$, then $(\operatorname{diag}(W^{(1)}, \dots, W^{(k)}), (B^{(1)}, \dots, B^{(k)})$ generates a zonotope in $\mathbb{R}^{n_1+\cdots+n_k}$ with points intersecting $\prod_i K_i$ even (odd) orthants of $\mathbb{R}^{n_1+\cdots+n_k}$.

Proposition 4.A.13. *Let $n, m \in \mathbb{N}$.*

- *If $4\lceil m/3 \rceil \leq n$, then $\mathrm{RBM}_{n,m} \cap \mathcal{H}_{n,2^m} \neq \emptyset$ and*

$$\mathrm{Mixt}^k(\mathcal{E}_n^1) \supseteq \mathrm{RBM}_{n,m} \quad \textit{if and only if} \quad k \geq 2^m \ .$$

- *If $4\lceil m/3 \rceil > n$, then*

$$\mathrm{Mixt}^k(\mathcal{E}_n^1) \supseteq \mathrm{RBM}_{n,m} \quad \textit{only if} \quad k \geq \min\{2^l + m - l, 2^{n-1}\} \ ,$$

  where $l := \max\{l \in \mathbb{N} \colon 4\lceil l/3 \rceil \leq n\}$.

Recall that $\mathcal{H}_{n,l}$ denotes the set of probability distributions on $\{0,1\}^n$ which have at least l strong modes (Definition 1.B.1).

Remark 4.A.14. The first item remains true if $m \mod (3)$=1 and $4\lfloor m/3 \rfloor + 2 \leq n$. Furthermore, if $n = 1, 2$, then $\mathrm{Mixt}^k(\overline{\mathcal{E}_n^1}) = \mathrm{RBM}_{n,k-1}$ for all k.

Proof. Let $4\lceil m/3 \rceil \leq n$. Consider the following parameters with $a, b \in \mathbb{R}$ and $W \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^n, C \in \mathbb{R}^m$:

$$W = \alpha \left(\begin{array}{ccccc|c} w & & & & & \\ & w & & & & \\ & & \ddots & & & 0 \\ & & & w & & \\ & & & & \tilde{w} & \end{array} \right) , \quad w = \begin{pmatrix} -1 & -1 & -1 & 1 \\ -1 & -1 & 1 & -1 \\ -1 & 1 & -1 & -1 \end{pmatrix} ,$$

where $\tilde{w}$ consists of the first or the first two rows of w. We use the following bias:

$$\begin{aligned} B =& \alpha \left(\ b \ \middle| \ b \ \middle| \ \cdots \ \middle| \ b \ \middle| \ (-1, \dots, -1) \ \right) , \\ b =& \frac{1}{2} \begin{pmatrix} 3 & 1 & 1 & 1 \end{pmatrix} , \\ C =& -W(1, \dots, 1)^\top = \alpha(2, \dots, 2)^\top . \end{aligned}$$

Denote by λ_i the set of entries $\{1,2,3,4\} + 4(i-1) \subset [n]$. For $\alpha \to \infty$ the joint distribution of the RBM satisfies:

$$p(v,h) \propto \begin{cases} 1 \, , & \text{if } \sum_{j \in \lambda_i} x_j \text{ is even for all } i \text{ and } x_j = 0 \text{ for all } j > 4\lceil m/3 \rceil \\ 0 \, , & \text{else} \end{cases} .$$

This implies that the visible probability distribution generated by the RBM is the uniform distribution with support on a subset of $Z_{+,n}$ of cardinality 2^m.

Now let $4\lceil m/3\rceil \geq n$. By the first part of this Proposition, the RBM with l hidden units can represent probability distributions with 2^l strong modes $\mathcal{C} \subseteq Z_{\pm,n}$. If p is in $\mathrm{RBM}_{n,l}$, then $\mathrm{RBM}_{n,l+1}$ contains $\alpha p+(1-\alpha)\delta_x$ for any arbitrary state x and $\alpha \in [0,1]$, see [72]. Hence each hidden unit additional to l allows us to increase the probability of any state in $Z_{\pm,n} \setminus \mathcal{C}$ while uniformly reducing the probability of all other states, and so increase the number of strong modes by one. □

Dimension of the RBM Model

Proposition 4.A.15.

$$\mathrm{RBM}_{n,m+1} = \{\mathrm{Mixt}(p, p * \overline{\mathcal{E}_n^1})\colon p \in \mathrm{RBM}_{n,m}\} \supseteq \mathrm{Mixt}(\mathrm{RBM}_{n,m} \setminus \partial\mathcal{P}_n, S(\mathcal{E}_n^1)) \ ,$$

where $p * q := \frac{p \cdot q}{\sum_x p(x)q(x)}$ *and* $S(\mathcal{E}_n^1) = \{p \in \overline{\mathcal{P}_n}\colon \mathrm{supp}(p) = \{x, y\}, d_H(x, y) \leq 1\}$ *consists of all faces of* $\mathcal{P}_n$ *which are contained in* $\overline{\mathcal{E}_n^1}$.

Proof. The first equality follows from eq. (4.8) in pg. 94. The inclusion relation follows from $p * S(\mathcal{E}_n^1) = S(\mathcal{E}_n^1)$ for all $p \in \mathcal{P}_n$. □

Corollary 4.A.16. $\dim(\mathrm{RBM}_{n,m})$ *strictly increases with* m *until reaching the value* $2^n - 1$.

Proof. If $\dim(\mathrm{RBM}_{n,m}) = d$, then there exists a subset of $\mathrm{RBM}_{n,m}$ which has a tangent space of dimension d. If a set $\mathcal{M} \subseteq \overline{\mathcal{P}}$ is closed under mixtures with δ_x for all x, then $\mathcal{M} = \overline{\mathcal{P}}$. For any $p \in \mathcal{P}$, any proper subset of $\{\delta_x - p\colon x \in \mathcal{X}\}$ consists of linearly independent vectors. The claim follows from the last inclusion in Proposition 4.A.15. □

Remark 4.A.17. In [31] it was shown that

$$\dim(\mathrm{RBM}_{n,m}) = \min\{nm + n + m, 2^n - 1\} \quad \text{when } m \leq 2^{n-\lceil \log_2(n+1)\rceil} \text{ or when } m \geq 2^{n-\lfloor \log_2(n+1)\rfloor} \ . \quad (4.21)$$

This formula holds if m is not too large, which is the case in most applications. If $(n-1)$ is not a power of two, the result doesn't hold for a number of values of m. In [31] it is conjectured that $\dim(\mathrm{RBM}_{n,m}) = \min\{nm + n + m, 2^n - 1\}$ for all n and m. Our Corollary 4.A.16 slightly extends the scope of the dimension formula (4.21).

4.B The Models $\mathrm{RBM}_{3,2}$ and $\mathrm{RBM}_{4,3}$

The Model $\mathrm{RBM}_{3,2}$

In Section 3 we showed that $\mathrm{RBM}_{n,m} = \overline{\mathcal{P}_n}$ only if $m \geq \lceil 2^n/(n+1)\rceil - 1$, and that if $m \geq 2^{n-1} - 1$, then $\mathrm{RBM}_{n,m} = \overline{\mathcal{P}_n}$. The RBM model with three visible units is the smallest example for which there is a gap between bounds for the sufficient and necessary number of hidden units of an RBM universal approximator. By [31] the dimension of $\mathrm{RBM}_{3,2}$ equals $2^n - 1$. Proposition 4.A.12 implies that $\mathrm{RBM}_{3,2}$ doesn't contain distributions with four strong modes (see also Appendix 3.B). In this section we take a closer look at the model from various

perspectives; the RBM joint model, its convex support and support sets, the KL-divergences to the $\mathrm{RBM}_{3,2}$, and the question how many probability distributions are not contained in $\mathrm{RBM}_{3,2}$?

By Theorem 4.2.1, $\mathrm{RBM}_{3,2}$ contains:

(i) Any mixture of an arbitrary product distribution and two further product distributions with disjoint supports.

(ii) Any mixture of two arbitrary product distributions and a further arbitrary distribution with support on a pair with Hamming distance one.

This includes any distribution with support of cardinality at most three, all distributions with support of cardinality four, except for those supported by $Z_{\pm,3}$, and any distribution with support contained in the union of three pairs with Hamming distance one.

For the complement of $\mathrm{RBM}_{3,2}$ we have:

Proposition 4.B.1. $\mathrm{RBM}_{3,2} \cap \mathcal{I}_3 = \emptyset$. *Furthermore,* $\mathcal{I}_3 \supsetneq \mathcal{H}_3$, $\mathrm{vol}(\mathcal{H}_3) = 0.0078 \cdot \mathrm{vol}(\mathcal{P}_3)$, *and* $\mathrm{ex}(\mathcal{H}_3^{\pm}) \subset \mathrm{RBM}_{3,2}$.

In particular, $\mathrm{RBM}_{3,m}$ is a universal approximator of distributions on $\{0,1\}^3$ if and only if $m \geq 3$.

Proof of Proposition 4.B.1. The first statement is a direct consequence of Proposition 4.A.12 and Proposition 4.A.7. More explicitly, if $\mathrm{RBM}_{3,2} \cap \mathcal{I}_3 \neq \emptyset$, then:

$$\mathrm{sgn}\begin{pmatrix} B \\ W_1 + B \\ W_2 + B \\ W_1 + W_2 + B \end{pmatrix} \overset{!}{=} \mathrm{rowperm}\,. \begin{pmatrix} + & - & - \\ - & + & - \\ - & - & + \\ + & + & + \end{pmatrix} . \tag{4.22}$$

This is a contradiction, because $(B) + (W_1 + W_2 + B) = (W_1 + B) + (W_2 + B)$ and the sign of the two addends in the left hand side of this expression must agree in at least one entry while on that same entry the two addends on the right hand side also must agree, but have the opposite sign. The analysis of the polytopes $\mathcal{H}_3^{\pm}$ is given below.

For three units and four strong modes we have:

$$\mathcal{H}_{3,4} = \mathcal{H}_3^{+} \cup \mathcal{H}_3^{-}\,, \quad \mathcal{H}_3^{\pm} := \bigcap_{z \in Z_{\pm}} \Big\{ p \in \overline{\mathcal{P}_3} : p(z) > \sum_{d_H(z,\hat{z})=1} p(\hat{z}) \Big\}\,. \tag{4.23}$$

By eq. (4.23), $\overline{\mathcal{H}_3^{+}}$ is the intersection of 8 half-spaces (this is a h-polytope) defined through the inequalities $p(z) \geq \sum_{y:d_H(z,y)=1} \ \forall z \in Z_+$ and $p(y) \geq 0 \ \forall y \in Z_-$. Using `Polymake` we find that $\mathcal{H}_3^{+}$ is a 7-dimensional simplex with volume $\mathrm{vol}(\mathcal{H}_3^{+}) = 0.0039 \cdot \mathrm{vol}(\mathcal{P}_3)$ and with vertices $\mathrm{ex}(\mathcal{H}_3^{+})$ given by the uniform probability distributions on the following sets:

$$\{000, 001, 011, 101\}, \{011, 101, 110, 111\}, \{000, 010, 011, 110\}, \{000, 100, 101, 110\},$$
$$\{000\}, \{011\}, \{101\}, \{110\} \tag{4.24}$$

All vertices of $\mathcal{H}_3^{+}$ are also vertices of $\mathcal{G}_3^{+}$. The vertices of $\mathcal{G}_3^{+}$ are listed in Table 1.1, page 32.

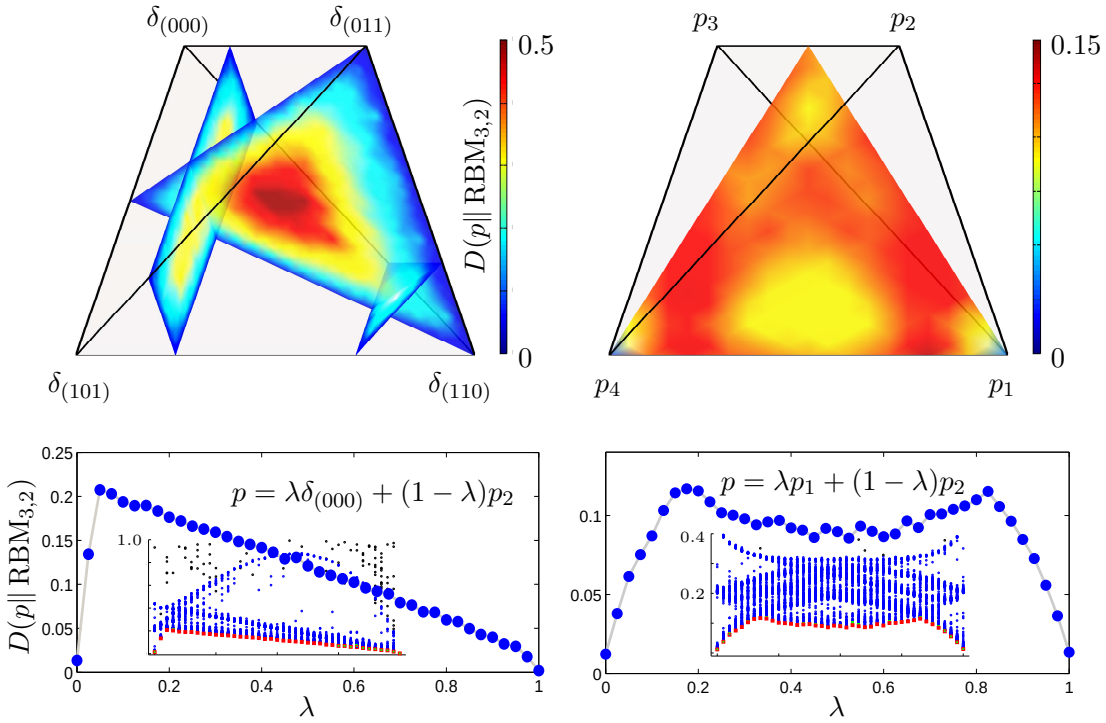

Figure 4.5: *Top left: This figure shows the face $\overline{\mathcal{P}}(Z_{+,3})$ of the probability simplex on $\{0,1\}^3$. The color indicates the KL-divergence to the best approximation that we found within $\mathrm{RBM}_{3,2}$ using contrastive divergence, maximum likelihood and extensive initializations of parameters (similar to the computations shown in Figure 4.3 from Section 4.2). The color is interpolated between a couple of hundreds of regularly spaced computed points. Top right: Similar computations for probability distributions contained in the simplex with vertices from the first group of extreme points of $\mathcal{H}_3^+$ listed in eq. (4.24). Bottom: Detail to the computations shown in the top. The figures show the KL-divergence from target distributions on the lines $\lambda\delta_{(000)}+(1-\lambda)p_2$ and $\lambda p_1+(1-\lambda)p_2$, which are edges of the polytope $\mathcal{H}_3^+$, to distributions within $\mathrm{RBM}_{3,2}$. The small inset figures show the results after training the RBM for each of the parameter initializations (the best approximations, shown in red, correspond to the large figures).*

The first four vertices listed in eq.(4.24) are mixtures of two point measures and one uniform distribution on a pair of Hamming distance one. The last four vertices are point measures. By Theorem 4.2.1, all vertices are contained in $\mathrm{RBM}_{3,2}$. The edges of the simplex $\mathcal{H}_3$ connecting two vertices from the second group are all in $\mathrm{RBM}_{3,2}$. The distributions in the relative interior of the edges of $\mathcal{H}_3$ between two vertices from the first group are in $\mathcal{I}_2^+$ and not in $\mathrm{RBM}_{3,2}$. At the lower right of Figure 4.5 we show the numerically computed KL-divergence from the points on such an edge to the model. The relative interior of edges connecting one vertex from the first group and one vertex from the second group are in $\mathrm{RBM}_{3,2}$ if they have support of cardinality four and they are not if they have support of cardinality five. Figure 4.5 bottom left shows the divergence from points on an edge of distributions with support of cardinality five to the model. □

The upper left corner of Figure 4.5 illustrates Proposition 4.B.1. It shows the simplex $\mathcal{P}(Z_{+,3})$ whose vertices are the probability distributions from the second group listed in eq. (4.24). The centroid is u_{Z_+}. Consider any distribution $q = \frac{1}{4}\delta_{z_1} + \frac{1}{4}\delta_{z_2} + \frac{1}{2}u_{\{z_3,z_4,y_1,y_2\}}$, with $\{z_i\}_{i=1}^4 =$

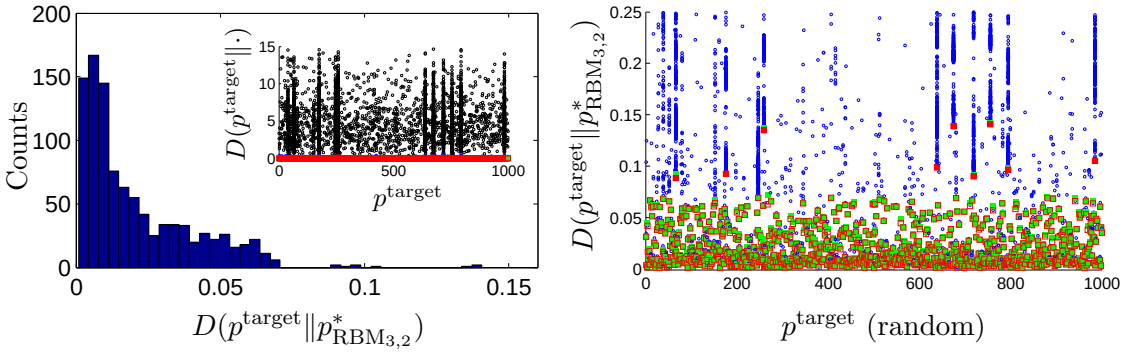

Figure 4.6: *Left: Histogram of KL-divergences from random targets to* $\mathrm{RBM}_{3,2}$ *after training. Inset and Right: KL-divergence from random targets to random probability distributions within* $\mathrm{RBM}_{3,2}$ *(black circles), to the probability distribution* $p^*_{\mathrm{RBM}_{3,2}}$ *that results from training* $\mathrm{RBM}_{3,2}$ *with data generated from the targets (blue circles). The best approximation of the targets after training is shown by the green squares and by the red empty squares after additional ML training. The red squares are filled if the final KL-divergence is larger than* 0.075.

$Z_\pm$ and $\{z_3, z_4, y_1, y_2\}$ a 2-face of C_3. The divergence from such a distribution to $u_{Z_\pm}$ is $D(u_{Z_\pm}\|q) = \frac{1}{2}$. Furthermore, such a q is contained in $\mathrm{RBM}_{3,2}$, because it is the mixture of two point measures and the uniform distribution on a face of C_3. In the upper right part of Figure 4.5 we show the simplex with vertices given by the first group of probability distributions: $p_1 = u_{\{(000),(001),(011),(101)\}}$, $p_2 = u_{\{(011),(101),(110),(111)\}}$, $p_3 = u_{\{(000),(010),(011),(110)\}}$, $p_4 = u_{\{(000),(100),(101),(110)\}}$.

Figure 4.6 shows the result of a computer aided examination of $D(p\|\,\mathrm{RBM}_{3,2})$ for many targets p. We recorded the largest KL-divergences for the following three targets (out of 1000 randomly generated):

$$\begin{array}{cccccccc}
(0.0298, & 0.1836, & 0.2325, & 0.0359, & 0.1232, & 0.0161, & 0.0409, & 0.3381) \\
(0.0936, & 0.2955, & 0.1452, & 0.0301, & 0.1959, & 0.0599, & 0.0059, & 0.1739) \\
(0.2170, & 0.1716, & 0.0236, & 0.2491, & 0.0083, & 0.1958, & 0.1185, & 0.0161) \\
\mathbf{(000)} & (001) & (010) & \mathbf{(011)} & (100) & \mathbf{(101)} & \mathbf{(110)} & (111)
\end{array}$$

The first two vectors are close to u_{Z_-}. The third vector is close to one of the edges described with eq. (4.24) (a probability distribution which is the convex combination of two vertices of $\mathcal{H}_3^\pm$ and has support of cardinality five). The mean KL-divergence from the targets to the trained RBM was 0.0208, the variance $3.8589 \cdot 10^{-4}$, the maximum 0.1406, and the minimum $8.0745 \cdot 10^{-4}$. Only 9 of 1000 random targets were farther than 0.075 to the trained RBM.

A Comparison of $\mathrm{RBM}_{3,2}$ and Mixture Models

We compare the KL-divergences from various distributions in $\mathcal{P}_3$ to *(i)* $\mathrm{RBM}_{3,2}$, *(ii)* $\mathrm{Mixt}^3(\mathcal{E}_3^1)$, *(iii)* the union of all mixtures of three product distributions with disjoint supports, and *(iv)* the union of all partition models with three blocks $\{\mathcal{P}_\xi \colon |\xi| = 3\}$. Figure 4.7 shows the result of our computations.

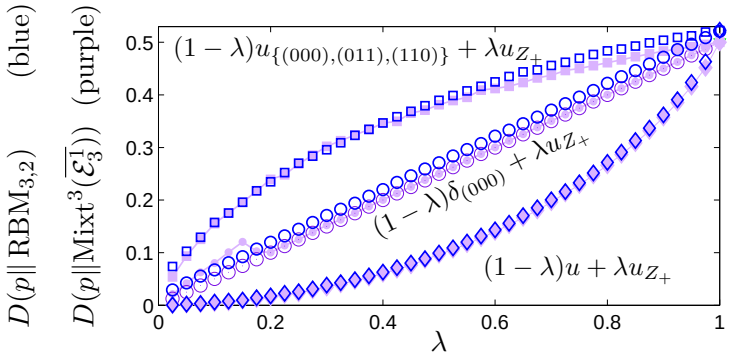

Figure 4.7: *Circles: KL-divergence from* $p = (1-\lambda)\delta_{(000)}+\lambda u_{Z_+}$, $0 \leq \lambda \leq 1$, *(the distributions in the line connecting a vertex to the centroid of the simplex* $\mathcal{P}(Z_+)$*), to* (i) $\mathrm{RBM}_{3,2}$ *(blue empty circles), computed using CD and ML,* (ii) $\mathrm{Mixt}^3(\mathcal{E}_3^1)$ *(purple filled circles), computed using EM methods,* (iii) *the union of partition models* $\cup_{\xi:|\xi=3|}\mathcal{P}_\xi$ *(purple empty circles), computed analytically. Squares: KL-divergence from* $(1-\lambda)u_{\{(000),(011),(110)\}}+\lambda u_{Z_+}$ *(the line connecting a centroid of a facet to the centroid of* $\mathcal{P}(Z_+)$*) to* $\mathrm{RBM}_{3,2}$ *(blue empty squares) and* $\mathrm{Mixt}^3(\mathcal{E}_3^1)$ *(purple filled squares). Diamonds: KL-divergence from* $(1-\lambda)u+\lambda u_{Z_+}$*, (distributions on the line from the uniform distribution to centroid of* $\mathcal{P}(Z_+)$*), to* $\mathrm{RBM}_{3,2}$ *(blue diamonds) and* $\mathrm{Mixt}^3(\mathcal{E}_3^1)$ *(purple filled diamonds). The KL-divergence was larger than zero whenever* $\lambda > 0$.

The divergences to the union of partition models with three blocks was computed analytically, (using Lagrange multipliers), on the line $p = (1-\lambda)\delta_{(000)}+\lambda u_{Z_+}$, $0 \leq \lambda \leq 1$. The solutions are of the form: $q = \alpha_1\delta_{(011)}+\alpha_2 u_{\{(110),(101),(100),(111)\}}+(1-\alpha_1-\alpha_2)\delta_{(000)}$, with mixture weights $\alpha_1 = -\frac{2}{\lambda}(1-\frac{1}{4}\lambda)\alpha_2+1$ and $\alpha_2 = (-\frac{4-2\lambda}{\lambda}+1)/(1-(4-2\lambda)(1-\frac{1}{4}\lambda)/(\frac{1}{2}\lambda^2))$. The resulting KL-divergence is a linear function of λ, see Figure 4.7.

For the model $\mathrm{Mixt}^3(\mathcal{E}_3^1)$ the KL-divergences are computed using a custom EM implementation. The divergence to this model was always positive (except for the target u). This reflects Proposition 1.B.15, which shows that $\mathrm{Mixt}^3(\mathcal{E}_3^1)$ contains no distributions with four modes. All the computed rI-projections to $\mathrm{Mixt}^3(\mathcal{E}_3^1)$ are in fact mixtures of three uniform distributions with disjoint supports (except for a few sub-optimal numerical results found for small values of λ). The projections into $\mathrm{RBM}_{3,2}$ were very similar. This suggests that the submodels of RBMs proposed in Theorem 4.2.1 contain the class of rI-projections of distributions with four modes into the RBM model.

The KL-divergence for targets in the line connecting the uniform distribution and the uniform distribution on Z_+ are always positive (except for the uniform distribution ($\lambda = 0$)), and virtually identical for all models.

The results from this section motivate the following conjecture:

Conjecture 4.B.2. *The model* $\mathrm{RBM}_{3,2}$ *doesn't contain any distribution with four modes. In particular, the uniform distribution is not an inner point of the model. The maximal KL-divergence to the model is* $\max_{p\in\overline{\mathcal{P}_3}}(p\|\,\mathrm{RBM}_{3,2}) = \frac{1}{2}$ *and there are exactly two maximizers; the uniform*

distributions supported by the perfect binary codes of minimum distance two Z_+ and Z_-.

The dimension of $\mathrm{RBM}_{3,2}$ and $\mathrm{Mixt}^3(\mathcal{E}_3^1)$ is seven. The model $\mathrm{RBM}_{3,2}$ contains many submodels of $\mathrm{Mixt}^3(\mathcal{E}_3^1)$, as explained in Theorem 4.2.1. It also contains mixtures of (four) elements in $\mathcal{E}_3^1$ with natural parameters in a two-dimensional subspace of $\mathbb{R}^3$. Such models don't contain any distribution of the form $\lambda u + (1-\lambda)u_{Z_{\pm,3}}$. See Corollary 1.B.20. By Proposition 4.B.1 and Proposition 1.B.15, the set of distributions with four strong modes is contained in the complement of both $\mathrm{RBM}_{3,2}$ and $\mathrm{Mixt}^3(\mathcal{E}_3^1)$. We showed in Proposition 4.A.13 that in general $\mathrm{Mixt}^{m+1}(\overline{\mathcal{E}_n^1}) \not\supseteq \mathrm{RBM}_{n,m}$. However, for the special case $m = 2$ and $n = 3$ we venture the following question: *Is the model* $\mathrm{RBM}_{3,2}$ *equal to* $\mathrm{Mixt}^3(\overline{\mathcal{E}_3^1})$*?*

Convex Support of $\mathcal{E}_{K_{3,2}}$

We used the computer software `Polymake` [47] to compute the f-vector of $\mathrm{cs}(\mathcal{E}_{K_{3,2}})$, (the number of proper faces of the polytope in each dimension, starting from dimension zero):

$$f(\mathrm{cs}(\mathcal{E}_{K_{3,2}})) = (32, 416, 2880, 10940, 24448, 33448, 28272, 14500, 4310, 684, 48)\ .$$

The polytope $\mathrm{cs}(\mathcal{E}_{K_{3,2}})$ is not two-neighborly, since in that case it would have $\binom{32}{2} = 496$ one-dimensional faces, instead of 416. In fact, the maximal dimension in which all faces are simplices is 1 (this also implies that the polytope is not two-neighborly, see [51, Theorem 7.4.3]), and the maximal dimension in which all faces are simple polytopes is 2. All vertices have degree 26 and every vertex is contained in 30 facets. The face lattice of this polytope fills a `11.2MB` plain text file.

There are 24 ways to get the 4 visible states of positive parity, $Z_{+,3}$, in combination with four different hidden states (in other words, there are 24 sets of binary vectors in $\{0,1\}^7$ of cardinality four which restrict to $Z_{+,3}$ on the first three entries and to four different vectors on the last two entries). Below we give one of them. The convex support $\mathrm{cs}(\mathcal{E}_{K_{3,2}})$ is symmetric with respect to relabeling of the states of the hidden nodes, and also with respect to permutation of nodes in the visible layer or in the hidden layer. These operations applied to one of the instances exhaust the 24 different cases:

$$\begin{array}{l} h_1 \\ h_2 \\ v_1 \\ v_2 \\ v_3 \end{array} \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix} =: (x^1, x^2, x^3, x^4)\ . \tag{4.25}$$

We can check whether these four vectors build a facial set or even an S-set of $\mathcal{E}_{K_{3,2}}$. If $\{x^i\}_{i=1}^4$ is an S-set, then the columns of the sufficient statistics matrix build a full rank matrix (see Proposition 1.A.3). The submatrix $T^{3,2}_{\{x^i\}_{i=1}^4}$ of the sufficient statistics of $\mathcal{E}_{K_{3,2}}$ (see eq. (4.18)) has full rank 4. However, $\{x^1, x^2, x^3, x^4\}$ is not a facial set of $\mathcal{E}_{K_{3,2}}$ for the following reason: We computed the vertices contained in the 48 facets of $\mathrm{cs}(\mathcal{E}_{K_{3,2}})$. Each facet has either 16 or 24 incident vertices. The smallest face (intersection of facets) containing the four vertices

$\{T^{3,2}_{x^i}\}_{i=1,2,3,4}$ is spanned by eight columns with the following indices:

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{pmatrix}. \tag{4.26}$$

This means that the set $\{x^1, \ldots, x^4\}$ from eq. (4.25) is not facial, as claimed. This gives an alternative proof of the fact that $\mathrm{RBM}_{3,2}$ can't represent any distribution with support $Z_{\pm,3}$, which we established in Proposition 4.B.1. The visible parts of the eight vectors in eq. (4.26) cover all $\{0,1\}^3$ (the subscripts give the decimal representation of the visible parts). This does not mean that any distribution from $\mathrm{RBM}_{3,2}$ taking positive values on $Z_\pm$ has full support on $\{0,1\}^3$, because we don't rule out facial sets for which two vectors have the same values on the hidden units and different values in $Z_\pm$ in the visible units.

The face of $\mathrm{cs}(\mathcal{E}_{K_{3,2}})$ described by the column vectors from eq. (4.26) is a three-neighborly simplicial polytope of dimension 6. It has 16 simplex facets of cardinality 6, which correspond to the following S-sets of $\mathcal{E}_{K_{3,2}}$:

$$\{012346\},\ \{012345\},\ \{012456\},\ \{123456\},\ \{234567\},\ \{134567\},\ \{123567\},\ \{024567\},$$
$$\{023457\},\ \{023467\},\ \{014567\},\ \{013457\},\ \{013467\},\ \{012567\},\ \{012357\},\ \{012367\}.$$

These sets represent precisely the S-sets of $\mathrm{RBM}_{3,2}$ which are the disjoint union of 3 pairs of vectors of with Hamming distance one (see Theorem 3.1.1).

The Model $\mathrm{RBM}_{4,3}$

The model $\mathrm{RBM}_{4,1}$ is a binary tree model and is equivalent to $\mathrm{Mixt}^2(\mathcal{E}^1_4)$ (see Theorem 4.2.1). The geometry of binary tree models was recently studied in [123]. The model $\mathrm{RBM}_{4,2}$ was studied in [32]. It has 14 parameters and codimension one in $\mathcal{P}_4$. In Corollary 1.B.17 we showed that $\mathrm{RBM}_{4,2}$ doesn't contain any distribution with eight modes. The model $\mathrm{RBM}_{4,3}$ is the smallest candidate of an RBM universal approximator with four visible units. This model has 19 parameters, while $\dim(\mathcal{P}_4) = 15$. On the other hand $(n-1)$, in this case 3, is a lower bound for the number of hidden units of a universal approximator on $\{0,1\}^n$. By Corollary 1.3.3, if the RBM represents a distribution with support $Z_{+,4}$, then each hidden state must produce a summand which is a point measure.

The computations shown in Figure 4.8 suggest that $\mathrm{RBM}_{4,3} = \overline{\mathcal{P}_4}$. We recorded a mean KL-divergence of 0.0359, a maximum value 0.1046, a minimum value 0.0064, and a corrected variance $\frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$ of $1.8458 \cdot 10^{-4}$. For comparison, in a similar experiment for $\mathrm{RBM}_{4,0}$ we recorded a mean KL-divergence 0.2054 and variance 0.0074. However, similar experiments for $\mathrm{RBM}_{3,2}$ also resulted in a low mean KL-divergence, and that model is not a universal approximator.

Proposition 4.B.3. *$\mathcal{E}_{K_{4,3}}$ has an S-set which restricts to $Z_{+,4}$ on the four visible entries. Hence* $\mathrm{RBM}_{4,3} \supset \overline{\mathcal{P}}(Z_{\pm,4})$.

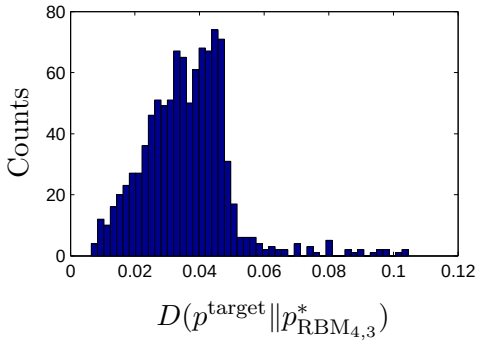

Figure 4.8: *Histogram of the KL-divergence from random target distributions to* $\mathrm{RBM}_{4,3}$ *after training. In this experiment we generated* 1000 *random target distributions in* $\mathcal{P}_4$*. Each target is the normalization of* 16 *uniformly distributed entries in* $[0, 1]$*. For each target,* $\mathrm{RBM}_{4,3}$ *was initialized at* 300 *different random parameters and trained on data generated from the target using contrastive divergence and ML methods. Whenever the KL-divergence was less than* 0.05*, the main training (and random initialization) was interrupted.*

Proof. The inclusion $u_{Z_+} \in \mathrm{RBM}_{4,3}$ follows from Proposition 4.A.13. It implies that there is a facial set $\mathcal{Y} \subseteq \{0,1\}^7$ of $\mathcal{E}_{K_{4,3}}$ with $\mathcal{Y}_V = Z_+$ and $\mathcal{Y}_H = \{0,1\}^3$. The columns of the sufficient statistics matrix $T^{4,3}$ of $\mathcal{E}_{K_{4,3}}$ corresponding to the state vectors within $\mathcal{Y}$ has full rank 8 (see the tabular in pg. 103). Hence $\mathcal{Y}$ is an S-set of $\mathcal{E}_{K_{4,3}}$ (see Lemma 1.2.5). This completes the proof. □

In particular, Proposition 4.B.3 emphasizes that the support sets of $p * \mathcal{E}$ for $p \in \mathrm{RBM}_{n,(m-1)}$ are *not* equal to the support sets of $\mathcal{E}$. The computation of $\mathcal{F}(\mathrm{cs}(\mathcal{E}_{K_{4,3}}))$ is quite expensive. We only computed the number of facets, which is 12480.

5 Model Design

In this chapter we are interested in the following quantity:

$$\max_{p\in\mathcal{G}} D(p\|\mathcal{M}) \ , \tag{5.1}$$

where $\mathcal{M} \subseteq \overline{\mathcal{P}}$ is a statistical model under consideration and $\mathcal{G} \subseteq \overline{\mathcal{P}}$ is a class of target probability distributions. An extensive evaluation of this function is an important and challenging problem. The case where $\mathcal{M}$ is an exponential family and $\mathcal{G} = \mathcal{P}$ was treated in [95]. In Chapter 1 and Chapter 3 we focused on universal approximation of probability distributions by mixtures of hierarchical models, RBMs and DBNs. The universal approximation problem is to reduce the expression (5.1) to zero for $\mathcal{G} = \mathcal{P}$ as a function of hyperparameters of $\mathcal{M}$ within some class of models. In Chapter 4 we studied the representational power of mixtures of independence models, RBMs, and DBNs which are not necessarily universal approximators and we estimated the approximation errors when approximating arbitrary probability distributions. This is, we estimated (5.1) for $\mathcal{G} = \mathcal{P}$ and a fixed $\mathcal{M}$ within the mixture, RBM, and DBN classes. This chapter provides a basis for even more extensive treatments of eq. (5.1).

In Section 5.1 we discuss a few examples as a proof of concept for the case where $\mathcal{G} \neq \mathcal{P}$. We discuss optimization problems for which there is always an optimizer within a specific region $\mathcal{G}$ of the search space, and we investigate the representation of deterministic kernels by RBMs. In Section 5.2 we treat the problem of reducing the number of parameters of $\mathcal{M}$, subject only to the condition that every point in $\mathcal{G}$ can be reached following a gradient, but not to the condition that $\mathcal{M}$ belongs to a particular class of models (e.g., the condition that $\mathcal{M}$ is represented by a stochastic network).

5.1 Restricted Boltzmann Machines and Deep Belief Networks

If $\mathcal{G} = \mathcal{E}^1_n$ and $\mathcal{M} = \mathrm{RBM}_{n,m}$ (or $\mathcal{M} = \mathrm{DBN}_{n,n_1,\ldots,n_l}$), the expression from eq. (5.1), $\max_{p\in\mathcal{G}} D(p\|\mathcal{M})$ has the trivial solution 0 for all m (and $n_1,\ldots,n_l$), because $\mathrm{RBM}_{n,0} = \overline{\mathcal{E}^1_n}$. The cases $\mathcal{G} = \mathcal{E}^2_n$ and $\mathcal{G} = \mathrm{Mixt}^2(\mathcal{E}^1_n)$ have an easy solution too: The model $\mathrm{Mixt}^2(\overline{\mathcal{E}^1_n})$ contains any mixture of two point measures and we know that the convex support of $\mathcal{E}^2_n$ is 3-neighborly (see Chapter 1). Therefore, $\overline{\mathcal{E}^2_n}$ and $\mathrm{Mixt}^2(\overline{\mathcal{E}^1_n})$ contain the probability distributions of the form $\frac{1}{2}(\delta_x + \delta_{1+x \,(\mathrm{mod}\, 2)})$, which are the global maximizers of $D(\cdot\|\mathcal{E}^1_n)$ within the full simplex $\overline{\mathcal{P}_n}$ (see Chapter 4). Therefore,

$$\max_{p\in\mathcal{G}} D(p\|\mathcal{E}^1_n) = (n-1) \qquad \text{for} \quad \mathcal{G} = \mathcal{E}^2_n \ \text{ and } \ \mathcal{G} = \mathrm{Mixt}^2(\mathcal{E}^1_n) \ . \tag{5.2}$$

By Theorem 4.2.1 $\mathrm{RBM}_{n,1} = \overline{\mathrm{Mixt}^2(\mathcal{E}^1_n)}$, and we get:

$$\max_{p\in\mathrm{Mixt}^2(\mathcal{E}^1)} D(p\|\,\mathrm{RBM}_{n,m}) = \begin{cases} (n-1), & \text{if } m = 0 \\ 0, & \text{if } m \geq 1 \end{cases} \ . \tag{5.3}$$

Distributions with support of bounded cardinality. An interesting case is when $\mathcal{G}$ contains only probability distributions with support of cardinality at most k for some $k \leq |\mathcal{X}|$. If $k = 1$, then the solution is trivial, because all point measures are product distributions. For some other k, we have

$$\max_{p \in \overline{\mathcal{P}_n} : |\mathrm{supp}(p)| \leq k} D(p \| \operatorname{RBM}_{n,m}) \leq \begin{cases} 0, \text{ for } k \leq (m+1) \text{ or } (m+1) \geq 2^{n-1} \\ \min\{(d_{k,n} - 1), D_{\operatorname{RBM}_{n,m}}\}, \text{ for } k > (m+1) \end{cases} \tag{5.4}$$

where $d_{k,n}$ is the maximal minimum distance of a binary code of length n and cardinality k, and $D_{\operatorname{RBM}_{n,m}}$ is bounded from above according to Theorem 4.2.2 (roughly by $(n-1) - \log(m+1)$).

Partition Models. Consider two partitions of $\mathcal{X}$

$$\xi = \{\mathcal{X}_1, \ldots, \mathcal{X}_m\} \quad \text{and} \quad \zeta = \{\mathcal{Y}_{1,1}, \ldots, \mathcal{Y}_{1,k}, \mathcal{Y}_{2,1}, \ldots, \mathcal{Y}_{m,k}\} \,,$$

such that all blocks of ξ have the same cardinality and each of them is the union of k blocks of ζ, $\mathcal{X}_i = \cup_{j=1}^{k} \mathcal{Y}_{i,j}$, with $\mathcal{Y}_{i,j}$ all of equal cardinality. Note that $\mathcal{P}_\zeta = \operatorname{Mixt}(\mathcal{P}_{\zeta_1}, \ldots, \mathcal{P}_{\zeta_m})$. By Lemma 4.1.1, we get the following:

$$\begin{aligned} \max_{p \in \mathcal{P}_\zeta} D(p \| \mathcal{P}_\xi) &= \max_{i=1,\ldots,m} \max_{p \in \mathcal{P}_{\zeta_i}} D(p \| u_{\mathcal{X}_i}) \\ &= \max_{p \in \mathcal{P}_{\zeta_i}} D(p \| u_{\mathcal{X}_i}) = D(u_{\mathcal{Y}_{1,1}} \| u_{\mathcal{X}_1}) = \log(k) \,. \end{aligned} \tag{5.5}$$

In fact, if ζ is a refinement of ξ, such that each block $\mathcal{X}_i$ of ξ is the union of k_i blocks $\{\mathcal{Y}_{j,1}, \ldots, \mathcal{Y}_{i,k_i}\}$ of ζ, then $\max_{p \in \mathcal{P}_\zeta} D(p \| \mathcal{P}_\xi) = \max_{i=1,\ldots,m, j=1,\ldots,k_i} \log(|\mathcal{X}_i| / |\mathcal{Y}_{i,j}|)$. If ξ consists of at most $(m+1)$ cubical sets, then this is an upper bound for $\max_{p \in \mathcal{P}_\zeta} D(p \| \operatorname{RBM}_{n,m})$.

Exchangeable Distributions. Consider now as target $\mathcal{G}$ the set of exchangeable distributions:

$$\mathcal{P}_{\text{exch},n} := \{p \in \overline{\mathcal{P}_n} \colon p(x) = p(y) \text{ whenever } \|x\|_1 = \|y\|_1\} \,. \tag{5.6}$$

This is the partition model whose blocks allocate the binary vectors of equal one norm $\|x\|_1 := \sum_{i=1}^{n} x_i$. We denote the corresponding partition by $\xi_{\text{exch},n}$. All vectors within a block of $\xi_{\text{exch},n}$ belong either to Z_+ or to Z_-. There are exactly $(n+1)$ blocks with cardinalities $\binom{n}{k}$ for $k = 0, \ldots, n$ and $\mathcal{P}_{\text{exch},n}$ is an n-simplex. In particular we have

$$\operatorname{Mixt}^m(\overline{\mathcal{E}_n^1}) \supset \mathcal{P}_{\text{exch},n} \quad \text{only if} \quad m \geq \binom{n}{\lfloor \frac{n}{2} \rfloor} \simeq 2^n / \sqrt{\tfrac{\pi}{2} n} \,. \tag{5.7}$$

The exact representation of all exchangeable distributions as mixtures of products requires almost as many components as the exact representation of all $\mathcal{P}_n$. This is because $\mathcal{P}_{\text{exch},n}$ contains distributions with a large support in Z_+. This should be compared to [36].

We now compare $\mathcal{P}_{\text{exch},n}$ with partition models contained in $\operatorname{RBM}_{n,m}$ and $\operatorname{DBN}(n_0^l)$. Let $\xi = \{\mathcal{X}_1, \ldots, \mathcal{X}_m\}$ be a partition model with m fixed cubical blocks. For any $p \in \mathcal{P}_{\text{exch},n}$, an rI-projection onto $\mathcal{P}_\xi$ is supported by the smallest union of blocks $\mathcal{X}_i$ which contains the support of p. Consider a $p \in \operatorname{ex} \mathcal{P}_{\text{exch},n}$, i.e., $p = u_C$ for some $C \in \xi_{\text{exch},n}$. The best approximation within $\mathcal{P}_\xi$ is $(u_C)_{\mathcal{P}_\xi} = \sum_{i \in [m]} \frac{|\mathcal{X}_i \cap C|}{|C|} u_{\mathcal{X}_i}$ and the KL-divergence is

$$D(u_c \| \mathcal{P}_\xi) = -H(u_C) + H\left((\tfrac{|\mathcal{X}_i \cap C|}{|C|})_{i=1,\ldots,m} \right) - \sum_{i \in [m]} \tfrac{|\mathcal{X}_i \cap C|}{|C|} \log \tfrac{1}{|\mathcal{X}_i|} \,, \tag{5.8}$$

where $H(p)$ denotes the Shannon entropy $H(p) := -\sum_x p(x)\log(p(x))$. This expression vanishes if *(i)* $C = \mathcal{X}_i$ for one i (in this case $C \cap \mathcal{X}_j = \emptyset$ for $j \neq i$), or if *(ii)* $m \geq |C|$ and C is covered by $|C|$ blocks of cardinality one. The case *(ii)* never occurs if $\mathcal{X}_i$ are cubical sets, since C has minimum distance two.

Input Output Maps. Given non-empty finite sets $\mathscr{X}$ and $\mathscr{Y}$, the stochastic matrices from $\mathscr{X}$ to $\mathscr{Y}$ are maps $(x, y) \mapsto \pi(x; y)$ satisfying

$$\pi(x; y) \geq 0 \text{ for all } x \in \mathscr{X}, y \in \mathscr{Y}, \quad \text{and}$$
$$\sum_{y \in \mathscr{Y}} \pi(x; y) = 1 \text{ for all } x \in \mathscr{X}.$$

The set of stochastic matrices is denoted by $\mathcal{C} := \mathcal{C}(\mathscr{X}; \mathscr{Y})$. Stochastic matrices are very general objects and can serve as models for individual neurons, neural networks, and policies. Each extreme point of this convex set corresponds to a deterministic function $g : \mathscr{X} \to \mathscr{Y}$ given as

$$\pi^{(g)}(x; y) = \begin{cases} 1, & \text{if } y = g(x) \\ 0, & \text{else} \end{cases} . \tag{5.9}$$

We consider the problem of maximizing an objective function $f : \mathcal{C} \to \mathbb{R}$ defined on the set $\mathcal{C}$ of stochastic matrices. A model $\mathcal{N} \subseteq \mathcal{C}$ is consistent with f, if the set of maximizers of f can be reached through the learning on $\mathcal{N}$. This implies that the maximizers of f should be contained in the closure of $\mathcal{N}$. If f is convex on $\mathcal{C}$, then each locally maximal value is attained at an extreme point of $\mathcal{C}$, and corresponds to a deterministic function. We refer to the following three examples in which optimal systems also turn out to be close to deterministic functions: *(i) Optimal policies in reinforcement learning* [111], *(ii) dynamics with maximal predictive information as considered in robotics* [121], and *(iii) dynamics of neural networks with maximal network information flow* [15]. This suggests to consider models that can approximate all extreme points of $\mathcal{C}$. In the following we concentrate on the first example to illustrate the main idea.

We divide the visible units of a network into an *input* and an *output* region, such that the conditional distributions of the output nodes given the state of the input nodes describe stochastic kernels. There are many ways of doing this. We restrict the discussion to dividing the visible layer of an RBM into two regions. B. Sallans and G. E. Hinton [100] used this ansatz for selecting good actions for Markov decision processes. Let n be the number of visible units, n_{in} the number of input units and n_{out} the number of output units. As outlined above, the deterministic kernels comprise the optimal solutions to various interesting problems. We are interested in the number of hidden units needed to represent the class of deterministic kernels. Let $v = (x, y)$ denote the state vector of all visible units, where $x \in \{0, 1\}^{n_{in}}$ and $y \in \{0, 1\}^{n_{out}}$. The deterministic kernels are represented by the probability distributions $p \in \overline{\mathcal{P}_n}$ for which the following holds:

$$p(\cdot|x) = \delta_y \quad \text{for some } y \in \{0, 1\}^{n_{out}} \quad \forall x \in \{0, 1\}^{n_{in}} . \tag{5.10}$$

There are $(2^{n_{in}})^{2^{n_{out}}}$ deterministic maps $\{0, 1\}^{n_{in}} \to \{0, 1\}^{n_{out}}; x \mapsto y_x$. Any particular function g is represented by any probability distribution contained in the following simplex:

$$\Delta_g := \Big\{ \sum_x \alpha(x)\delta_{(x,y_x)} \colon \sum_x \alpha(x) = 1,\ \alpha(x) > 0\ \forall x \Big\} \subset \overline{\mathcal{P}_n} . \tag{5.11}$$

The RBM model can represent any g if and only if it intersects all open simplices of the form given in eq. (5.11). The set Δ_g consists of all probability distributions $p \in \overline{\mathcal{P}_n}$ strictly supported by a binary code which has exactly one element in each cylinder set $\{(x,y) \in \{0,1\}^n \colon y \in \{0,1\}^{n_{out}}\}\ \forall x \in \{0,1\}^{n_{in}}$. Denote $\xi_{n_{in},n_{out}}$ this collection of cylinder sets. Hence, the representation of deterministic kernels is a special case of representation of probability distributions with support of cardinality bounded by $2^{n_{in}}$. Any element from Δ_g is a mixture of $2^{n_{in}}$ point measures. Since $\mathrm{RBM}_{n,m}$ contains any mixture of $(m+1)$ product distributions with disjoint supports (see Theorem 4.2.1), it can represent any deterministic function whenever $(m+1) \geq 2^{n_{in}}$. In this case, however, each function is considerably overparameterized. In fact, $\mathrm{RBM}_{n,m} \supset \Delta_g\ \forall g$, and the fiber of each g has dimension at least $2^{n_{in}} - 1$. This is just an upper bound on the minimal number of parameters, but at the same time:

Whenever $1 \leq n_{in} \leq n-1$, there is a deterministic kernel which is represented by the probability distributions supported by a binary code of minimum distance at least two, because each block in the partition $\xi_{n_{in},n_{out}}$ intersects $Z_{+,n}$. We can state the following:

Proposition 5.1.1. *Let $n \in \mathbb{N}$ and $1 \leq n_{in} \leq n-1$. The model $\mathrm{Mixt}^m(\overline{\mathcal{E}_n^1})$ can represent all deterministic kernels if and only if $m \geq 2^{n_{in}}$. If $(m+1) \geq 2^{n_{in}}$, then $\mathrm{RBM}_{n,m}$ can represent all deterministic kernels. If the conditions on m are satisfied, then for either model the fiber of each deterministic kernel has dimension larger or equal to $2^{n_{in}} - 1$.*

This result should be compared to the results form Appendix 4.A. The number of hidden units of an RBM can be potentially reduced for specific classes of deterministic kernels, even if n_{in} is as large as $n-1$. For example the parity function $f \colon (x_1, \ldots, x_{n-1}) \mapsto \sum_{i \in [n-1]} x_i \mod 2$ is represented by distributions supported on $Z_{\pm,n}$, which can be contained in $\mathrm{RBM}_{n,m}$ for $m \ll 2^{n-1} - 1$ depending on n.

5.2 An Approach to Reduce the Parameter Space of Learning Systems

We propose ways of defining models of stochastic matrices that are compatible with the maximization of expected reward in reinforcement learning theory. Our approach is based on information geometry and aims at the reduction of model parameters as a way to improve gradient learning processes. We present two-dimensional models which contain all extreme points from the set of stochastic matrices, and which allow a simple implementation of *natural gradient* methods.

5.2.1 Geometric Idea

We first consider general convex sets and return to stochastic matrices in Section 5.2.2. The convex hull of a finite set $\xi^{(1)}, \ldots, \xi^{(n)}$ in $\mathbb{R}^d$ is

$$\mathcal{C} := \Big\{ \sum_{i=1}^{n} p(i)\, \xi^{(i)} \ : \ p(i) \geq 0\ \forall i \text{ and } \sum_{i=1}^{n} p(i) = 1 \Big\} \,. \tag{5.12}$$

The set of extreme points of this polytope $\mathcal{C}$ is a subset of $\{\xi^{(1)}, \ldots, \xi^{(n)}\}$. In general, there are many ways to represent a point $x \in \mathcal{C}$ as a convex combination of the extreme points. Here, we

are interested in convex combinations obtained from an exponential family. To be more precise, denote $\mathcal{P}$ the set of probability measures $p = (p(1), \dots, p(n)) \in \mathbb{R}^n$ and consider the map

$$\mathsf{m}: \ \mathcal{P} \ \to \ \mathcal{C}, \quad p \ \mapsto \ \sum_{i=1}^{n} p(i)\, \xi^{(i)} \ .$$

Consider an exponential family $\mathcal{E}_\phi$ with sufficient statistics $\phi = (\phi_1, \dots, \phi_l)$ on $\{1, \dots, n\}$. Denote $\mathcal{C}_\phi$ the image of $\mathcal{E}_\phi$ by the map m. With the choice $\phi_k^*(i) := \xi_k^{(i)}$ for $i = 1, \dots, n$ and $k = 1, \dots, d$, the closure of $\mathcal{E}_\phi$ can be identified with the polytope $\mathcal{C}$ (see Section 1.1). This allows to define natural geometric structures on $\mathcal{C}$, such as a Fisher metric, by using the natural structures on the simplex $\mathcal{P}$. In the context of stochastic matrices this leads to a Fisher metric that has been studied by Lebanon [74] based on an approach by Čencov. The above construction also motivates the following definition: We call a family $\mathcal{C}_\phi$ an *exponential family* in $\mathcal{C}$ if the vectors ϕ_k, $k = 1, \dots, l$, are contained in the linear span of the vectors ϕ_k^*, $k = 1, \dots, d$.

In general, the families $\mathcal{C}_\phi$ are not exponential families but projections of exponential families. We are mainly interested in models that are compatible with the maximization of a given function $f : \mathcal{C} \to \mathbb{R}$ in the sense that the closure of $\mathcal{C}_\phi$ should contain the maximizers of f. This is clearly not the only consistency condition, but here we focus on this assumption only.

As stated above, in many cases the local maximizers of f are elements of the set $\{\xi^{(1)}, \dots, \xi^{(n)}\}$, and hence the problem stated above reduces to finding a family $\phi = (\phi_1, \dots, \phi_l)$ of functions such that $\mathcal{C}_\phi$ contains that set in its closure. This is always possible with only two functions ϕ_1, ϕ_2. One such family can be constructed as follows: Consider a one-to-one map φ of the n points $\xi^{(1)}, \dots, \xi^{(n)}$ into $\mathbb{R}$, for instance $\xi^{(i)} \mapsto i$, $i = 1, \dots, n$, and the following family of distributions:

$$p_{\alpha,\beta}(i) = \frac{e^{-\beta\left(\varphi(\xi^{(i)})-\alpha\right)^2}}{\sum_{j=1}^{n} e^{-\beta\left(\varphi(\xi^{(j)})-\alpha\right)^2}} = \frac{e^{\lambda_1\, \phi_1(i) + \lambda_2\, \phi_2(i)}}{\sum_{j=1}^{n} e^{\lambda_1\, \phi_1(j) + \lambda_2\, \phi_2(j)}} \ , \tag{5.13}$$

where $\phi_1(i) := \varphi(\xi^{(i)})$, $\phi_2(i) := \varphi^2(\xi^{(i)})$, and $\lambda_1 := 2\,\alpha\,\beta$, $\lambda_2 := -\beta$. It is easy to see that for $\alpha = \varphi(\xi^{(i)})$ and $\beta \to \infty$, the distribution $p_{\alpha,\beta}$ converges to the point measure concentrated in i. The convex combination $\sum_{j=1}^{n} p_{\alpha,\beta}(i)\, \xi^{(i)}$ therefore converges to the point $\xi^{(i)}$. This proves that the closure of this two-dimensional family in $\mathcal{C}$ contains all the points $\xi^{(i)}$, $i = 1, \dots, n$. In general, the geometric properties of this family strongly depend on φ, as we discuss in the following section.

Figure 5.1 shows the set $\mathcal{C}_\phi$ for the choice $\phi(i) = \begin{pmatrix} \sin(2\pi i/n) \\ \cos(2\pi i/n) \end{pmatrix}$ and $n = 4$ (left), $n = 8$ (center) and $n = 16$ (right). In this case, the convex support of the exponential family $\mathcal{E}_\phi$ is a regular polygon. The left part shows $\{\xi^{(i)}\} = \{\delta_i\}_{i=1}^4$, for which $\mathcal{C}_\phi = \mathcal{E}_\phi \subseteq \mathcal{P}_n$. For the given choice of ϕ, $\mathcal{E}_\phi$ corresponds precisely to the independence model of two binary variables. For the figure in the center we set $\{\xi^{(i)}\}_{i=1}^8 = \{0,1\}^3$. In the right figure we set $\{\xi^{(i)}\}_{i=1}^{16} = \{0,1\}^4$. In this case $\mathcal{C}$ is isomorphic to the set of $n \times 2$ stochastic matrices (see eq. (5.15)).

Remark 5.2.1. (Hamiltonian exponential families). In the remainder of this section we will only need the above introduced two-dimensional exponential families. However, we want to discuss a

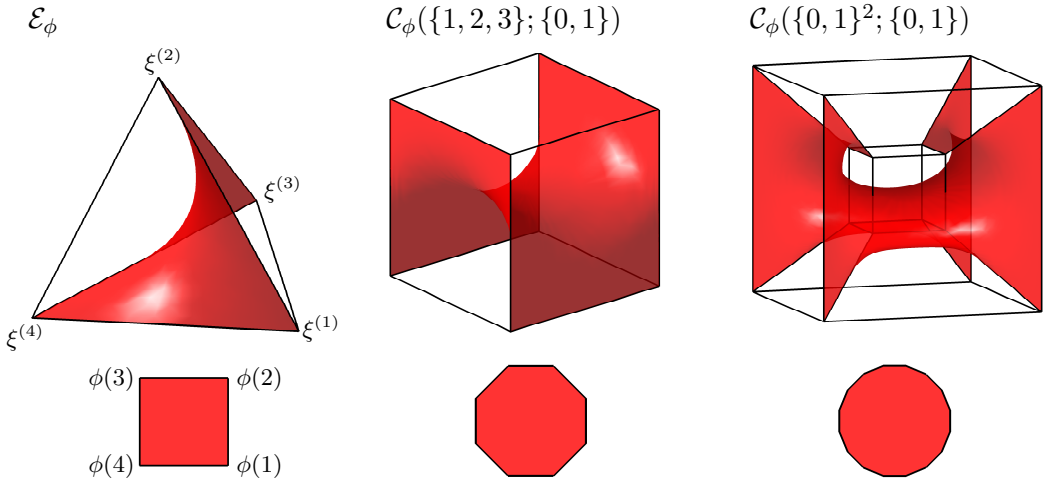

Figure 5.1: *Two-dimensional models approaching all vertices of various polytopes. The models are defined as in eq.* (5.13) *for various choices of* n, $\{\xi(i)\}_{i=1}^n$, *and* ϕ. *We use two-dimensional exponential families which have a regular polygon as convex support. The left model corresponds to an exponential family in the probability simplex* $\mathcal{P}(\{1,2,3,4\})$ *and corresponds to the independence model of two binary variables. The figures in the center and in the right correspond to two-dimensional families of stochastic matrices with three, respectively four inputs and two outputs. The figures in the bottom show the convex support of the exponential family* $\mathcal{E}_\phi$, $\mathrm{cs}(\mathcal{E}_\phi) = \mathrm{conv}\{\phi(i)\}_{i=1}^n$.

natural extension of the idea: We can use low-dimensional exponential families which approach all probability distributions of support of cardinality κ. The smallest such exponential family on $\mathcal{X}$ has a convex support which is lowest dimensional among all κ-neighborly polytopes with $|\mathcal{X}|$ vertices. For the cyclic polytope $C(d,n)$ every k vertices determine a $(k-1)$-face for all $k \leq \frac{d}{2}$, and its f-vector (containing the number of faces in each dimension) satisfies $f_i(C(v,d)) = \binom{v}{i+1}$ for $0 \leq i \leq \lfloor \frac{1}{2}d \rfloor$. The remaining f_i are completely determined by the *Dehn-Sommerville equations* [51, Theorem 4.7.1, Theorem 9.2.1]. The *Upper bound theorem* states that cyclic polytopes have, among all convex d-polytopes with n vertices, the largest number of faces in each dimension, see [81, 82]. The polytope $C(n,d)$ is realized as the convex hull of $\{x(t_i) \in \mathbb{R}^d\}$, where $t_1 < \cdots < t_n$, and x is the *moment map* $x(t) = (t^1, \ldots, t^d)$. Hence, the following sufficient statistics defines a κ-neighborly exponential family $\mathcal{E}$ of dimension 2κ.

$$\phi_k(i) = t_i^k \quad i = 1, \ldots, n, \ k = 1, \ldots, d\,. \tag{5.14}$$

The family from eq. (5.13) fits in this framework, since a two-dimensional exponential family approaching all point measures has a convex support which is a polygon, and any polygon is combinatorially equivalent to a cyclic polytope (any n different points on $(t, t^2), t \in \mathbb{R}$ are the vertices of an n-gon). See Figure 5.1. An additional comment: Combinatorially equivalent polytopes do not necessarily induce the same exponential family (this is the case for affinely equivalent polytopes).

5.2.2 An Application to Reward Maximization

Although the number of extreme points of the set of stochastic matrices $\mathcal{C}(\mathscr{X};\mathscr{Y})$ is $|\mathscr{Y}|^{|\mathscr{X}|}$, according to Section 5.2.1 there always exists a two-dimensional manifold that reaches all of

them. Note that in the particular case of N binary neurons we have $\mathscr{X} = \mathscr{Y} = \{0,1\}^N$ and therefore $(2^N)^{(2^N)}$ extreme points.

To illustrate the geometric idea we consider the example $\mathscr{X} = \{1,2,3\}$ and $\mathscr{Y} = \{1,2\}$. This can, for instance, serve as a model for policies with three states and two actions. In this case $\mathcal{C}$ is a subset of $\mathbb{R}^{\mathscr{X}\times\mathscr{Y}} \cong \mathbb{R}^6$ which can be identified with the hypercube $[0,1]^3$ through the following parametrization (see Figure 5.2 A):

$$[0,1]^3 \ni (r,s,t) \;\mapsto\; \begin{pmatrix} r & 1-r \\ s & 1-s \\ t & 1-t \end{pmatrix} . \tag{5.15}$$

To test the properties of that family with respect to the optimization of a function, we consider a map $(s,a) \mapsto \mathcal{R}^a_s$, which we interpret as *reward* that an agent receives if it performs action a after having seen state s. The policy of the agent is described by a stochastic matrix $\pi(s;a)$. The expected reward can be written as

$$f(\pi) = \sum_s p^\pi(s) \sum_a \pi(s;a)\, \mathcal{R}^a_s \,. \tag{5.16}$$

In reinforcement learning, there are several choices of p^π (see [112]). Here we simplify our study by assuming p^π to be the uniform measure.

We investigate the influence of the map φ and compare the natural gradient flow (gradient with respect to the Fisher metric, see [5]) with the ordinary gradient. For the experiments we drew a random reward matrix $\mathcal{R}$ and applied gradient ascent (with fixed step size) on $f(\pi)$ restricted to our model and several choices of φ (see Figures 5.2 A/B for typical outcomes). The optimization results strongly depend on φ. We restricted ourselves to the case that φ maps the vertices of $\mathcal{C}$ onto the numbers $\{1,\dots,n\}$. Such a map is equivalent to an ordering of the vertices. We recorded the best results when φ corresponds to a Hamiltonian cycle on the graph of the polytope $\mathcal{C}$, i.e., a closed path on the edges of the polytope's graph that visits each vertex exactly once and returns to the starting vertex. This way φ preserves the locality in $\mathcal{C}$, and the resulting model $\mathcal{C}_\phi$ is a smooth manifold. In Figure 5.2 A, both methods reach the global optimum $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$. In Figure 5.2 B, φ is 'unordered'. In this case the landscape $f(\pi_{\alpha,\beta})$ is more intricate and contains several local maxima. The natural gradient method only converged to a local but not global optimum, and the ordinary gradient method failed.

A *reflected Gray code* is defined recursively as follows (see [27]): *(i)* Take the $n-1$ bit Gray code (a list of binary vectors of length $(n-1)$), reflect the list (first vector becomes last vector). *(ii)* Set the prefix 0 to the original list, and 1 to the reflected list. *(iii)* Concatenate the two prefixed lists.

Proposition 5.2.2. *Let $|\mathscr{X}| = n$, $|\mathscr{Y}| = 2$, and let $f(\pi) = \sum_s \sum_a \pi(s;a)\mathcal{R}^a_s$. If φ enumerates $\mathrm{ex}(\mathcal{C}(\mathscr{X},\mathscr{Y}))$ according to an n-bit reflected Gray code and $\mathcal{R}$ is a random, generic reward matrix, then the uniform stochastic matrix $\pi \propto 1$ is not a local optimizer of $f|_{\mathcal{C}_\phi}$, and the gradient points in the direction of the global maximizer.*

Proof. See pg. 129. □

Figure 5.3 illustrates Proposition 5.2.2 for two small families.

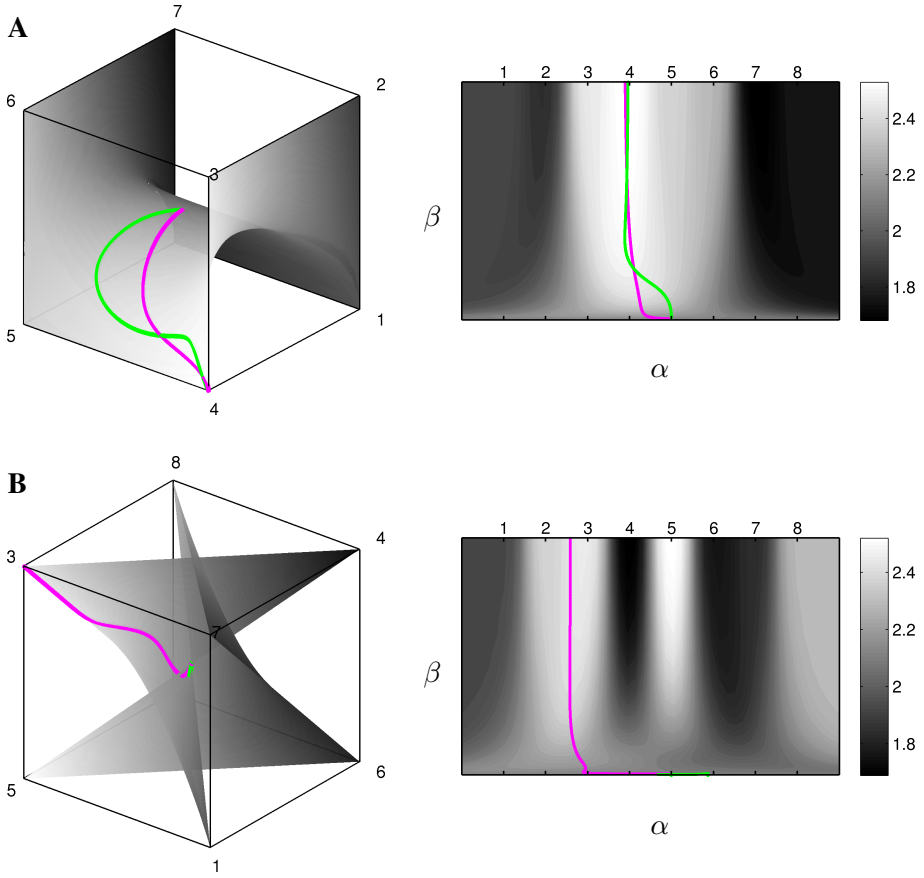

Figure 5.2: *Optimization with ordinary (green learning curves) and natural (magenta learning curves) gradient on the model $\mathcal{C}_\phi$ for two different choices of φ. Each vertex ξ of the cube is labeled by $\varphi(\xi)$. A: A Hamiltonian cycle $\varphi = (1,2,3,4,5,6,7,8)$. B: An arbitrary map $\varphi = (1,7,3,5,2,8,4,6)$.*

5.2.3 A Construction of Neuromanifolds

Here we approach implementations of policies π in the context of neural networks. We start with the case of two binary input neurons and one binary output neuron (Figure 5.4, left). All neurons are considered to be binary with values 0 and 1. The input-output mapping is modelled in terms of a stochastic matrix π. The set of such 4×2-matrices forms a four-dimensional cube. A prominent neuronal model assumes synaptic weights w_1 and w_2 assigned to the directed edges and a bias b. The probability for the output 1, which corresponds to the spiking of the neuron, is then given as

$$\pi(x_1, x_2; 1) \quad = \quad \frac{1}{1 + e^{-(w_1 x_1 + w_2 x_2 - b)}}. \tag{5.17}$$

This defines a three-dimensional model in the four-dimensional cube, see Figure 5.5. Some extreme points are not contained in this model, e.g. the matrix $\pi(0,0;1) = \pi(1,1;1) = 0$, $\pi(0,1;1) = \pi(1,0;1) = 1$. This corresponds to the well-known fact that the standard model

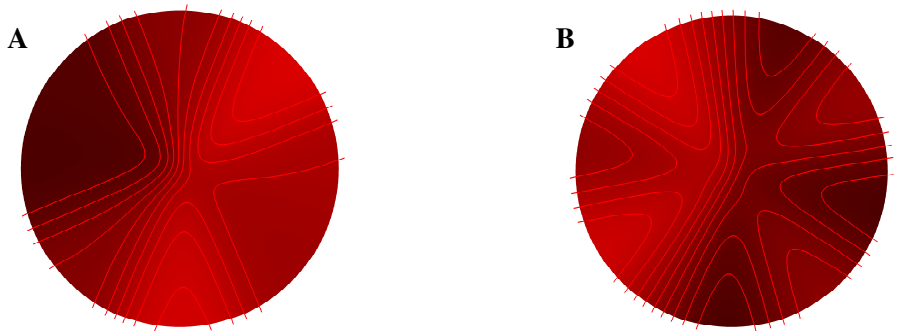

Figure 5.3: *This figure illustrates Proposition 5.2.2. Shown are the level surfaces of the expected reward for a random reward matrix $\mathcal{R}_s^a$ (see eq. (5.16)) as a function of the natural parameters of the two-dimensional exponential family $\mathcal{E}_\phi$ for systems with two outputs and three (A), respectively four (B) inputs. Here the enumeration φ of the deterministic functions corresponds to a reflected (cyclic) binary Gray code. The corresponding models of stochastic matrices $\mathcal{C}_\phi$ are shown in Figure 5.1 (center and right). The center of the depicted region of the parameter space corresponds to the uniform matrix $\pi(s,a) = \frac{1}{2}\ \forall s, a$.*

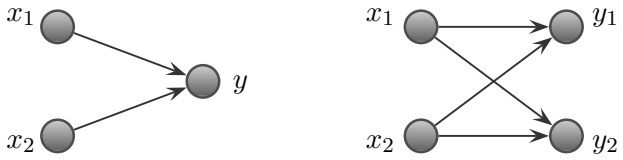

Figure 5.4: *Two simple neural networks.*

cannot represent the XOR-function. On the other hand, it is possible to reach all extreme points, including the XOR-function, with the two-dimensional models introduced previously in this section. However, there are various drawbacks of our models in comparison with the standard model. They are not exponential families but only projections. Moreover, we do not have a neurophysiological interpretation of the parameters.

We now discuss models for the case of one additional output neuron. The system is modelled by stochastic 4×4 matrices, which form the 12-dimensional polytope $\mathcal{C} := \mathcal{C}(\{0,1\}^2; \{0,1\}^2)$. A natural assumption is the independence of the outputs Y_1 and Y_2 given the input pair X_1, X_2. This is the case if and only if the input-output map of each neuron i is modelled by a separate stochastic matrix π_i, $i = 1, 2$. The stochastic matrix of the whole system is given by

$$\pi(x_1, x_2; y_1, y_2) \ = \ \pi_1(x_1, x_2; y_1) \cdot \pi_2(x_1, x_2; y_2).$$

This defines an 8-dimensional model $\mathcal{N}_{\text{product}}$ that contains all extreme points of $\mathcal{C}$. Furthermore, it contains the submodel $\mathcal{N}_{\text{standard}}$ given by the additional requirement that π_1 and π_2 are of the form (5.17). The model $\mathcal{N}_{\text{standard}}$ is an exponential family of dimension 6. However, as in the one-neuron case, $\mathcal{N}_{\text{standard}}$ does not reach all extreme points. Another submodel $\mathcal{N}_{\text{new}}$ of $\mathcal{N}_{\text{product}}$ is defined by modelling each of the stochastic matrices π_i in terms of two parameters as described above. The following table gives a synopsis:

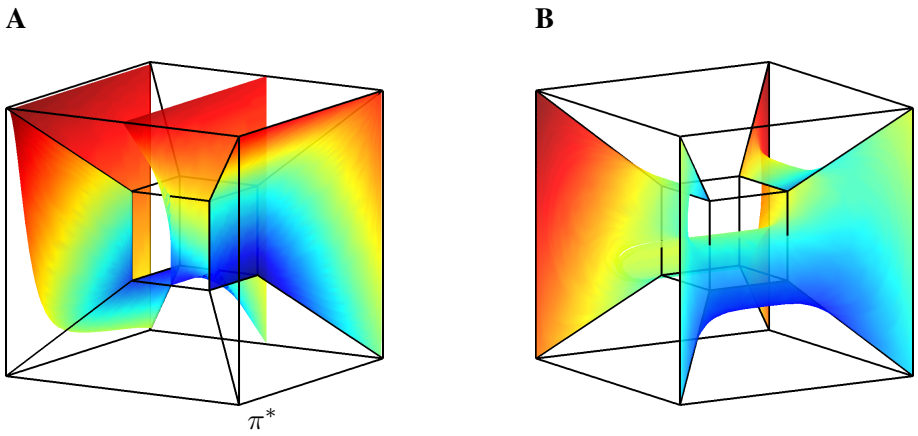

Figure 5.5: *A: The standard model given in eq.* (5.17) *for three values of the bias parameter b. The deterministic policy π^* is not contained in this model. B: The new model introduced in Section 5.2.1. The color indicates the value of the long-term expected reward for a randomly chosen $\mathcal{R}$.*

model	dim.	exponential family	reaches all extreme points
$\mathcal{C}$	12	yes	yes
$\mathcal{N}_{\text{product}}$	8	yes	yes
$\mathcal{N}_{\text{standard}}$	6	yes	no
$\mathcal{N}_{\text{new}}$	4	no	yes

We conclude this section with the maximization of a reward function in the family $\mathcal{N}_{\text{new}}$, as in the previous section. Figure 5.6 shows a histogram of the results for a fixed randomly chosen reward $\mathcal{R}$ after 500 steps for ordinary gradient and natural gradient methods. We chose a constant learning rate and 5,000 different random initial values. Both methods find 3 local maxima. The natural gradient process tends to converge faster. Furthermore, it finds the global maximum for a majority of the initial values, which is not the case for the ordinary gradient.

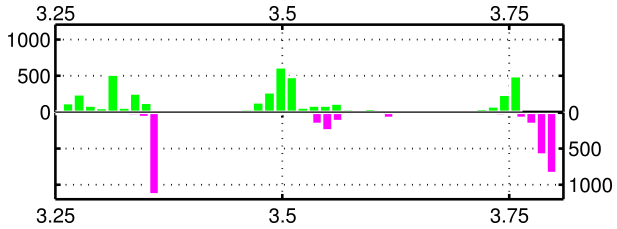

Figure 5.6: *Histogram of the objective value $f(\pi)$ after 500 steps of gradient ascent in $\mathcal{N}_{new}$. Magenta: natural gradient. Green: ordinary gradient.*

5.A Proofs and Details

Fisher Information and Parameter Updates

This appendix contains details to the computations presented in the Sections 5.2.2 and 5.2.3.

Two-dimensional Models

We consider the set of stochastic matrices which can be written as convex combinations of a set of extreme stochastic matrices $\{\xi^{(i)}\}_{i=1}^n$, whereas the mixture weights are given by probability distributions from a two dimensional model $\{p_{\beta,\alpha} \in \mathcal{P}_n \colon \beta, \alpha \in \mathbb{R}\} : \pi_{\beta,\alpha} = \sum_\xi p_{\beta,\alpha}(\xi)\xi$. For the mixture weights we take the following two dimensional set:

$$p_{\beta,\alpha}(\xi) = \frac{1}{Z} \exp(\beta \cos(\alpha - \varphi(\xi))) \;, \tag{5.18}$$

where $\varphi(\xi) = 2\pi \frac{k_\xi}{n}$, $\{k_\xi\} = \{0, \ldots, n-1\}$, is an enumeration of the extreme points. In this parametrization, α has the interpretation of an angle and β corresponds to the *inverse temperature* as used in *stochastic relaxation* and in the Gibbs-Boltzmann distribution of statistical physics. Note that this model corresponds to the exponential family $\mathcal{E}_\phi$ with sufficient statistics $\phi_1 = \cos(\varphi)$, $\phi_2 = \sin(\varphi)$, and sometimes it is more convenient to use the natural parameters of the sufficient statistics ϕ_1 and ϕ_2.

The derivatives of the log-probability from eq. (5.18) are:

$$\begin{aligned} \nabla \log p_{\beta,\alpha}(\xi) &= (\partial_\beta, \partial_\alpha) \log p_{\beta,\alpha}(\xi) \\ &= \begin{pmatrix} \cos(\alpha - \varphi(\xi)) - \mathbb{E}_{\beta,\alpha}[\cos(\alpha - \varphi)] \\ -\beta \sin(\alpha - \varphi(\xi)) + \mathbb{E}_{\beta,\alpha}[\beta \sin(\alpha - \varphi)] \end{pmatrix}^\top , \end{aligned}$$

where $\mathbb{E}_{\beta,\alpha}[f] := \sum_\xi p_{\beta,\alpha}(\xi) f(\xi)$. Clearly we also have $\nabla p_{\beta,\alpha}(\xi) = (\nabla \log p_{\beta,\alpha}(\xi)) p_{\alpha,\beta}(\xi)$. We abbreviate $p_{\beta,\alpha}$ by p and $\mathbb{E}_{\beta,\alpha}$ by $\mathbb{E}$. The Fisher information matrix is given by:

$$G(\beta,\alpha) = \begin{pmatrix} \mathbb{E}[\partial_\beta(\log p)\, \partial_\beta(\log p)] & \mathbb{E}[\partial_\beta(\log p)\, \partial_\alpha(\log p)] \\ \mathbb{E}[\partial_\beta(\log p)\, \partial_\alpha) \log p)] & \mathbb{E}[\partial_\alpha(\log p)\, \partial_\alpha(\log p)] \end{pmatrix} =: \begin{pmatrix} g_{11} & g_{12} \\ g_{12} & g_{22} \end{pmatrix} ,$$

where $g_{ij} = g_{ij}(\beta, \alpha)$. In this case (two parameters), the inverse matrix can be written explicitly:

$$G^{-1} = \begin{pmatrix} g_{22} & -g_{12} \\ -g_{12} & g_{11} \end{pmatrix} (g_{11} g_{22} - g_{12} g_{12})^{-1} \;.$$

The gradient of π is given by $\nabla \pi = \sum_\xi (\nabla p(\xi)) \xi$, and the derivative of the objective (simplified long-term expected reward for a reward matrix $\mathcal{R} \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{Y}|}$) is:

$$\nabla \rho^\pi = \sum_s \sum_a (\nabla \pi(s,a)) \mathcal{R}(s,a) \;. \tag{5.19}$$

The parameter updates are $(\beta, \alpha) \leftarrow (\beta, \alpha) + \eta \widetilde{\nabla} \rho^\pi$, where for the *natural gradient* we set $\widetilde{\nabla} = G^{-1} \nabla$, and for the ordinary gradient we set $\widetilde{\nabla} = \nabla$.

Product Model for Multiple Output Units

We now discuss the product model $\mathcal{N}_{\text{new}}$ for an action space of cardinality larger than two, as the stochastic maps arising from a feedforward network with several output neurons. For clarity we focus on the case of two binary inputs and two binary outputs. The model is given by product distributions of the form: $p(a_1, a_2 | s_1, s_2) = p(a_1 | s_1, s_2) \cdot p(a_2 | s_1, s_2)$, and stochastic matrices

of the form: $\pi = \pi_1 \cdot \pi_2 \in \mathcal{C}(\mathcal{X};\mathcal{Y}) \subset \mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}|}$, where $\mathcal{Y} = \mathcal{Y}_1 \times \mathcal{Y}_2 = \{0,1\} \times \{0,1\}$, $\mathcal{X} = \{s_1, s_2\} = \{0,1\}^2$, and

$$\pi_l = \sum_{i=1}^{|\mathcal{Y}_l|^{|\mathcal{X}|}} \mathrm{ex}_{\mathcal{C}_l}^{(i)}\, p_l(i), \quad \text{for} \quad l = 1,2\ , \tag{5.20}$$

where $\mathcal{C}_l := \mathcal{C}(\mathcal{X};\mathcal{Y}_l)$ and $\mathrm{ex}_{\mathcal{C}_l}^{(i)} \in \{0,1\}^{|\mathcal{X}|\times 2}$ denotes the i-th extreme point of $\mathcal{C}_l$ (given some enumeration). The probability distribution p_l belongs to a two-dimensional model in $\mathcal{P}(|\mathcal{Y}_l|^{|\mathcal{X}|})$; for example the one defined in eq. (5.13):

$$p_l(i) = \frac{1}{Z}\exp(-\lambda_1^l(\varphi(i) - \lambda_2^l)^2)\ , \tag{5.21}$$

and where we can set $\varphi(i) = i - 1$. The extreme stochastic matrices $\mathrm{ex}_{\mathcal{C}_l}^{(i)}$ correspond to deterministic binary functions with a single output. A sensible enumeration is the following:

$$\mathrm{ex}_{\mathcal{C}_l}^{(i)} = (G_i^\top, 1 - G_i^\top)\ , \tag{5.22}$$

where G_i are the rows of a cyclic Gray code, (each G_i is a binary vector of length $|\mathcal{X}|$, and subsequent rows differ in exactly one entry). This enumeration defines a Hamiltonian cycle on the graph of the $|\mathcal{X}|$-dimensional unit cube and intends to preserve locality (for $\mathrm{ex}_{\mathcal{C}_l}^{(i)}$ and $\mathrm{ex}_{\mathcal{C}_l}^{(i+1)}$ are *near* by each other). Finally

$$\pi = \pi_1 \cdot \pi_2 = \sum_{i,j=1}^{|\mathcal{Y}_1|^{|\mathcal{X}|}} \left(\mathrm{ex}_{\mathcal{C}_1}^{(i)} \tilde{\otimes}\, \mathrm{ex}_{\mathcal{C}_2}^{(j)}\right) p_1(i) \cdot p_2(j)\ , \tag{5.23}$$

where for $A \in \mathbb{R}^{n\times m}$, and $B \in \mathbb{R}^{n\times k}$ with rows A_i and B_i we write $A\tilde{\otimes}B$ for the matrix with rows $A_i \otimes B_i$, (i.e., $\tilde{\otimes}$ is a row-wise Kronecker product). Note that

$$\{\mathrm{ex}_{\mathcal{C}_1}^{(i)} \tilde{\otimes}\, \mathrm{ex}_{\mathcal{C}_2}^{(j)}\}_{i,j=1}^{|\mathcal{Y}_1|^{|\mathcal{X}|}} = \mathrm{ex}_{\mathcal{C}(\mathcal{X};\mathcal{Y}_1\times\mathcal{Y}_2)}\ , \tag{5.24}$$

and the elements of $\mathrm{ex}_{\mathcal{C}(\mathcal{X};\mathcal{Y})}$ can be indexed by the tuple (i,j).

The ordinary gradient with respect to the given parametrization is:

$$\begin{aligned}
\nabla p(i,j) &= \left(\partial_{\lambda_1^1}, \partial_{\lambda_2^1}, \partial_{\lambda_1^2}, \partial_{\lambda_2^2}\right) p(i,j) \\
&= \begin{pmatrix}
(-(\varphi(i) - \lambda_2^1)^2 p_1(i) + p_1(i)(p_1((\varphi - \lambda_2^1)^2)'))p_2(j) \\
(2\lambda_1^1(\varphi(i) - \lambda_2^1)p_1(i) - p_1(i)(p_1(2\lambda_1^1(\varphi - \lambda_2^1))'))p_2(j) \\
(-(\varphi(j) - \lambda_2^2)^2 p_2(j) + p_2(j)(p_2((\varphi - \lambda_2^2)^2)'))p_1(i) \\
(2\lambda_1^2(\varphi(j) - \lambda_2^2)p_2(j) - p_2(j)(p_2(2\lambda_1^2(\varphi - \lambda_2^2))'))p(i)
\end{pmatrix}
\end{aligned}$$

For the policy matrix we have $\sum_{ij} \nabla p(i,j)\, \mathrm{ex}_{\mathcal{C}}^{(i,j)} = (\partial_{\lambda_1^1}\pi, \partial_{\lambda_2^1}\pi, \partial_{\lambda_1^2}\pi, \partial_{\lambda_2^2}\pi)$, and the derivative of the long-term expected reward is:

$$\nabla\rho^\pi = \left(\sum_s \sum_a \partial_\lambda \pi(s,a)\mathcal{R}(s,a)\right)_{\lambda=\lambda_1^1,\lambda_2^1,\lambda_1^2,\lambda_2^2}$$

The Fisher information matrix is given by $g_{\alpha\beta} = \sum_{i,j=1} p(i,j)\partial_\alpha \log p(i,j)\partial_\beta \log p(i,j)$, where

$$\begin{aligned}
\partial_{\lambda_1^1} \log p(i,j) &= -(\varphi(i) - \lambda_2^1)^2 + (p_1((\varphi - \lambda_2^1)^2)') \\
\partial_{\lambda_2^1} \log p(i,j) &= 2\lambda_1^2(\varphi(i) - \lambda_2^1) - (p_1(2\lambda_1^1(\varphi - \lambda_2^1))') \\
\partial_{\lambda_1^2} \log p(i,j) &= -(\varphi(j) - \lambda_2^2)^2 + (p_2((\varphi - \lambda_2^2)^2)') \\
\partial_{\lambda_2^2} \log p(i,j) &= 2\lambda_1^2(\varphi(j) - \lambda_2^2) - (p_2(2\lambda_1^2(\varphi - \lambda_2^2))')
\end{aligned}$$

The parameter updates are then given as $\lambda \leftarrow \lambda + \eta\tilde{\nabla}\rho^\pi$, where for the natural gradient we set $\tilde{\nabla} = g^{-1}\nabla$, and η is the *learning rate*.

Gradient and Global Optimizers

We consider the two-dimensional model $\mathcal{C}_\phi = \pi_{\beta,\alpha} = \sum_\xi p_{\beta,\alpha}\xi$, where $\{\xi\}$ are the extreme points of the polytope of stochastic matrices $\mathcal{C}$ and $\mathcal{E}_\phi = \{p_{\beta,\alpha}\}$ is the two-dimensional exponential family from eq. (5.18). If the extreme point $\tilde{\xi}$ is the global optimizer of $\rho^\pi = \sum\sum \pi\mathcal{R}$, does ρ^{π_t} strictly increase (not decrease) along the trajectory

$$\pi_t = \sum_\xi \frac{1}{Z}\exp(t\cos(\varphi(\tilde{\xi}) - \varphi(\xi)))\xi \ , \quad t = [0,\infty) \tag{5.25}$$

until reaching its global maximum $\rho^{\tilde{\xi}}$? Note that if $\tilde{\xi}$ is the optimum for ρ^π, then the entries with value 1 in each row of $\tilde{\xi}$ are at the same places as the maxima of $\mathcal{R}$ in each row, and hence: If the velocity of (5.25) is a vector in the interior (closure) of the orthant with signs $\mathrm{sgn}(2\tilde{\xi} - 1)$, then ρ^{π_t} strictly increases (does not decrease) as t increases.

We show that the sign condition given above is satisfied at the origin, in the case that $\mathcal{C}$ are $n \times 2$ stochastic matrices and $\varphi(\xi)$ numbers the extreme matrices (2^n binary vectors of length n) along a reflected Gray code, which is the content of Proposition 5.2.2.

Proof of Proposition 5.2.2. We simplify the notation of the extreme $n\times 2$ matrices: $\xi \mapsto (2\xi_{i,1} - 1)_{i=1,\ldots,n}$ (we consider only the first column, and replace zeros by negative ones). For the velocity we have the following:

$$\begin{aligned}
\partial_t \pi_t &= \partial_t \sum_\xi \frac{1}{Z}\exp(t\cos(\varphi(\tilde{\xi}) - \varphi(\xi)))\xi \\
&= \sum_\xi \Big(\cos(\varphi(\tilde{\xi}) - \varphi(\xi)) - \mathbb{E}_{t,\varphi(\tilde{\xi})}[\cos(\varphi(\tilde{\xi}) - \varphi(\xi))]\Big)\, p_{t,\varphi(\tilde{\xi})}\xi \ .
\end{aligned}$$

At time $t = 0$ this expression is simplified to: $\sum_\xi \cos(\varphi(\tilde{\xi}) - \varphi(\xi))\frac{1}{2^n}\xi$, and we need to show that:

$$\left(\sum_\xi \cos(\varphi(\tilde{\xi}) - \varphi(\xi))(\xi)_i\right)(\tilde{\xi})_i \geq 0 \ . \tag{5.26}$$

Consider a reflected n-bit Gray code (we replace the zeros by minus ones): $G = (G_{x,i})_{x,i} \in \{-1,1\}^{2^n\times n}$. Denote the x-th row by $G_{x,:}$ and the i-th column by $G_{:,i}$. We introduce the abbreviation $(c_{\tilde{x}})_x := (\cos(\frac{2\pi}{2^n}(k_x - k_{\tilde{x}})))_x$, and rewrite (5.26) for arbitrary $\tilde{\xi}$ as:

$$\langle c_{\tilde{x}}, G_{:,i}\rangle\, G_{\tilde{x},i} \geq 0 \quad \forall\tilde{x}, \forall i \ . \tag{5.27}$$

For $i < n$ the i-th column is:

$$G_{:,i} = (\underbrace{-1-1}_{2^{i-1}}\underbrace{\underbrace{11}_{2^i}\underbrace{-1-1}_{2^i}\ldots\underbrace{11}_{2^i}}_{2^{n-i}-1}\underbrace{-1-1}_{2^{i-1}})^\top ,$$

and for $i = n$ it is $G_{:,n} = (\underbrace{-1,\ldots,-1}_{2^{n-1}},\underbrace{1,\ldots,1}_{2^{n-1}})^\top$.

We see that for any i and an arbitrary $\tilde{x}$, the column $G_{:,i}$ is a cyclic concatenation of blocks of length 2^i (or 2^{n-1} for $i = n$) with alternating sign. If $2^i < 2^{n-1}$, then (5.27) vanishes for all $\tilde{x}$. This is because the frequency of $G_{:,i}$ is a multiple of the frequency of $c_{\tilde{x}}$. If $2^i = 2^{n-1}$, then (5.27) is strictly greater than 0 for all $\tilde{x}$. This is because $(c_{\tilde{x}})_{x=\tilde{x}} = 1$, such that for at least half of all x the signs of $c_{\tilde{x}}$ and $G_{:,i}$ are equal, and in particular for $\tilde{x}$, where $c_{\tilde{x}}$ attains its maximum. This is the case for exactly two columns of G, namely $i = n$ and $i = n-1$. This completes the proof. □

Outlook

Mixtures of discrete exponential families. Our results from Chapter 1 show that the combinatorics of support sets of exponential families provide a powerful tool to assess the expressive power of mixture models. This approach enabled us to compute sharp bounds on the number of mixtures of exponential families which can represent an arbitrary target distribution. The computation of support sets of hierarchical models is intimately related to the computation of *Markov bases* of graphical models, which is a challenging problem of today in algebraic statistics, see [38, 35, 40, 48, 65, 64, 66, 96]. Our approach poses the following problem: *What is the minimal number of simplex faces of the convex support of a hierarchical model which suffices to cover all vertices?* We showed this covering number is equal to the Carathéodory number of independence models. We showed that this is not true for general exponential families, even if they contain all point measures in their closures (see Proposition 1.2.8). Notwithstanding, we conjecture that the statement holds for hierarchical models more general than the independence models.

Our analysis of mixtures of k-interaction models shows that it is possible to control the representational power of stochastic networks involving higher-order interactions. The idea of using higher-order interactions in learning systems is not new [102], but it attracts new interest in the community, as new learning methods are being developed along with specific network design for predetermined tasks, like the restricted three-way interaction Boltzmann Machines [84, 83].

One of the ideas presented in this thesis is to study mixture models using a restriction to mixtures of models with disjoint supports. This allows a considerable simplification of the analysis. This approach allows us to assess the maximum Kullback-Leibler divergence of mixture models and to find maximum likelihood projections into the mixtures using the existing framework for exponential families [95] and coding theory. In Chapters 1 and 4 we made the following observation: *If $\mathcal{E}$ denotes an independence model, $\mathrm{Mixt}^m(\mathcal{E}) = \mathcal{P}$ if and only if $\mathrm{Mixt}^m(\partial\mathcal{E}) = \overline{\mathcal{P}}$, and furthermore, $\max_{p\in\mathcal{P}} D(p\|\mathrm{Mixt}^m(\mathcal{E})) = \max_{p\in\mathcal{P}} D(p\|\{\mathrm{Mixt}(\mathcal{E}_1,\ldots,\mathcal{E}_k)\colon k \leq m$ and $\mathcal{E}_i \subset \overline{\mathcal{E}}\})$ for various values of m, where $\mathcal{E}_i$ are supported by disjoint facial sets of $\mathcal{E}$.* A more detailed analysis of this circumstance and eventual generalizations of our results can result in even more powerful tools to treat models with latent variables.

We demonstrated that the analysis of the modes of mixture models is a good way to recognize and describe the complement of mixture models. The further elaboration of the ideas presented in Appendix 1.B is a promising research direction.

Convex Subsets, Secants, Geodesics and Convex Hulls. The geometry of exponential families has been studied for a long time, resulting in beautiful mathematical results and important implications to applied fields [16, 4, 9, 113, 6, 7]. The study of mixtures of exponential families poses a great number of challenging questions about the geometry of exponential families which demand yet further research. In Chapter 2 we posed a number of questions motivated by the problem of computing inclusion relations of mixture models of exponential families. We

developed a series of tools to attack these questions. In particular, we trust that our results on secants of exponential families and convex subsets of exponential families find applications in identifiability of parameters of mixtures of exponential families, computation of volumes and dimensions of mixture models, and finally in estimation of model approximation errors. For example, there are closed form solutions of $D_{\mathcal{E}}$ for convex exponential families. Many interesting continuations of our ideas from Chapter 2 are conceivable. Interesting extensions include the treatment of intersections of exponential families and α-families of dimension $d \geq 2$, and more extensive treatment of our analysis of α-mixtures of exponential families, especially in the case of strictly positive basis points.

Universal Approximation Results for RBMs and DBNs. In this thesis we provided new upper bounds for the minimal number of parameters of universal approximators of type RBM and DBN. These bounds show striking similitudes to the sharp bounds that we computed for mixtures of product distributions. Our method exploits the ansatz of probability sharing proposed in [73] exhaustively and therefore, an approach based only on similar ideas will unlikely allow for improvements. Although our bounds are sharp for small models, we still don't know if our results represent the minimal size of DBN and RBM universal approximators. Establishing the minimal size of universal approximators would be a major theoretical contribution to this research field.

Expressive Power and Approximation Errors of RBMs and DBNs. We related the expressive power of RBMs and DBNs to mixtures of product distributions and unions of partition models. This picture of the models allows us to compute important quantities, such as the model approximation errors. We provided for the first time an account on the KL-approximation errors of these models. We bounded the approximation errors from above in terms of the number of hidden units and hidden layers of the systems. These results represent a significant advance in the task of model selection and assessment of the performance of learning algorithms. A desirable continuation of our results would be to find the maximal representatives from the proposed classes of submodels, which are contained in $\mathrm{RBM}_{n,m}$, respectively $\mathrm{DBN}_{n,n_1,\ldots,n_l}$. In particular: *What is the largest k for which* $\mathrm{Mixt}^k(\mathcal{E}^1_{n,\mathrm{bin}}) \subseteq \mathrm{RBM}_{n,m}$*?* An important contribution and natural extension of our results, would be to bound the maximal KL-divergence for RBMs and DBNs from below. Clearly, this relates to the problem of finding the smallest RBM and DBN approximators discussed above and is a difficult problem. Nevertheless, we computed lower bounds for the maximal KL-approximation errors of mixtures of independence models, and give thereby a starting point for treating more complicated mixtures.

We believe that our approximation error bounds for narrow DBNs can be improved through a more detailed analysis of the proposed submodels. A more fine-grained analysis of the proposed submodels, especially for DBNs containing only few hidden layers of different widths (which are most common in practice) would be an immediate and promising continuation of our work. Another interesting extension would be the treatment of higher-order machines and other deep architectures used in machine learning.

Model design for RBM*s and* DBN*s.* The problem of model design is most relevant for applications. This problem is inherently related to the problems of universal approximation and expressive power of statistical models. The results contained in this thesis represent a substantial advance in this direction. At the same time, model design entails considerable challenges;

including the assessment of lower bounds on the approximation errors, as discussed in Chapter 4, and the problem of mathematically describing a set of candidate target distributions. We showed that our approach to the representational power of mixture models, RBMs, and DBNs qualifies to attack general problems of model design. However, a more extensive treatment of the problem remains to be accomplished in future work.

In Section 5.2 we showed that it is possible to define low-dimensional learning systems which are guaranteed to find the solutions of diverse optimization problems. This approach is related to the realizability of neighborly polytopes. A specific problem arising from our considerations is the following: *Find polytopes which are as regular as possible and combinatorially similar to cyclic polytopes.* A cyclic polytope usually is much more elongated along one direction than the others. For example, the polygons realized as the convex hull of points $\{(t_i, t_i^2)\}_{i\in I}$ are not regular. As a consequence, the exponential families defined through such polytopes are "asymmetric". This can have an impact on applications. The two dimensional case is solved by the regular polygons. However, in three dimensions there are only 5 regular polyhedra (the Platonic solids), and in dimension larger than four, the only regular polytopes are the cross polytope, the simplex and the hypercube [30].

A related problem is the following: *Find an exponential family with the smallest possible dimension for which* $\mathrm{Mixt}^m(\overline{\mathcal{E}}) \supseteq \cup_{\mathcal{Y}\subseteq\mathcal{X}:|\mathcal{Y}|\leq\kappa}\overline{\mathcal{P}}(\mathcal{Y})$. Note that this problem is not necessarily solved by the smallest dimensional $\lceil\frac{\kappa}{m}\rceil$-neighborly exponential families. In particular, a k-neighborly polytope is $(2k-1)$-simplicial. If any κ vertices can be covered by m (2^k-1)-faces, and $k < \lceil\frac{\kappa}{m}\rceil$, then this would be a better solution. This can be compared with the S-set coverings for hierarchical models computed in Section 1.

Model Dimensions. From the perspective of algebraic statistics, the dimension and identifiability of parameters in statistical models are most fundamental problems. The dimension of the RBM model was treated in [31], revealing that this model has the expected dimension in many cases. Nevertheless, still for certain combinations of the number of visible and hidden units, the dimension remains unknown. We provided a slight extension of the result from [31] using Proposition 4.A.15. This proposition can be used to derive an RBM version of Terracini's lemma, which is a standard tool for estimating the dimension of secant varieties. Evaluating this approach is subject of our current research.

The dimension of the submodels of DBNs that we proposed in this work yield a lower bound on the dimension of the DBN model. These submodels do not target the estimation of the model dimension and are not likely to be optimal. However, currently there are no other results on the dimension of DBN models. The problem of computing the dimension of deep learning systems is a major target of current mathematical research. Our results can be used as starting point to assess the dimension of DBN models.

Singular model selection. Singular models are statistical models with a parametrization which is non-trivially degenerated. Models within a nested hierarchy are likely to be singular. The *learning coefficients* of singular models, in the sense of S. Wantanabe [119], can be used in a form of *extended Bayes Information Criterion* and provide a means to compare the learning performance of different models. A target of current mathematical research on RBMs and DBNs is to compute the corresponding learning coefficients. M. Aoyagi [11] studies the learning coefficients of

special RBMs. An analysis of DBNs still remains to be accomplished. The learning coefficients are functions of the fibers of the target distributions, and therefore, of the target distributions themselves. The computation of the learning coefficients represents a challenging problem in this emerging field of research. In to-date considerations, the target distributions are mostly assumed to be contained in the models. A treatment of asymptotic likelihood integrals which accounts for the expressive power of the models under consideration represents a challenging problem for future research.

Bibliography

[1] H. Abo, G. Ottaviani, and C. Peterson, *Induction for secant varieties of segre varieties*, Transactions of the American Mathematical Society **361** (2009), 767–792.

[2] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, *A learning algorithm for Boltzmann Machines*, Cognitive Science **9** (1985), no. 1, 147–169.

[3] N. Alon, *On the rigidity of Hadamard matrices*, unpublished, 1990.

[4] S. Amari, *Differential-geometrical methods in statistics*, Lecture Notes in Statistics, vol. 28, Springer-Verlag, New York, 1985.

[5] ———, *Natural gradient works efficiently in learning*, Neural Computation **10** (1998), no. 2, 251–276.

[6] ———, *Information geometry on hierarchy of probability distributions*, IEEE Trans. Inf. Theory **47** (2001), no. 5, 1701–1711.

[7] ———, *Conditional mixture model for correlated neural spikes*, Neural Computation **22** (2010), 1718–1736.

[8] S. Amari, O. E. Barndorff-Nielsen, R. E. Kaas, S. L. Lauritzen, and C. R. Rao, *Differential geometry in statistical inference*, Lecture Notes Monograph Ser., vol. 10, Inst. Math. Statistics, Hayward California, Hayward, California, 1987.

[9] S. Amari, K. Kurata, and H. Nagaoka, *Information geometry of Boltzmann machines*, IEEE Tran. Neural Netw. **3** (1992), no. 2, 260–271.

[10] S. Amari and H. Nagaoka, *Methods of information geometry*, Translations of Mathematical Monographs, vol. 191, American Mathematical Society, 2007.

[11] M. Aoyagi, *Stochastic complexity and generalization error of a Restricted Boltzmann Machine in Bayesian estimation*, J. Mach. Learn. Res. **99** (2010), 1243–1272.

[12] N. Ay, *An information-geometric approach to a theory of pragmatic structuring*, The Annals of Probability **30** (2002), no. 1, pp. 416–436 (English).

[13] N. Ay and A. Knauf, *Maximizing multi-information*, Kybernetika **42** (2006), 517–538.

[14] N. Ay, G. F. Montúfar, and J. Rauh, *Selection criteria for neuromanifolds of stochastic dynamics*, Advances in Cognitive Neurodynamics (III), Springer, 2011.

[15] N. Ay and T. Wennekers, *Dynamical properties of strongly interacting Markov chains*, Neural Networks **16** (2003), 1483–1497.

[16] O. Barndorff-Nielsen, *Information and exponential families in statisical theory*, Wiley series in probability and mathematical statistics. Probability and mathematical statistics, J. Wiley, 1979.

[17] Y. Bengio, *Learning deep architectures for AI*, Foundations and Trends in Machine Learning **2** (2009), no. 1, 1–127.

[18] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, *Greedy layer-wise training of deep networks*, Advances in Neural Information Processing Systems 19 (NIPS'06) (Bernhard Schölkopf, John Platt, and Thomas Hoffman, eds.), MIT Press, 2007, pp. 153–160.

[19] Y. Bengio and Y. LeCun, *Scaling learning algorithms towards AI*, Large Scale Kernel Machines (L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, eds.), MIT Press, 2007.

[20] C. M. Bishop, *Pattern recognition and machine learning (information science and statistics)*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[21] A. Björner, M. Las Vergnas, B. Sturmfels, N White, and G. M. Ziegler, *Oriented matroids*, Encyclopedia of Mathematics and Its Applications, vol. 46, Cambridge University Press, 1999.

[22] C. Bocci and L. Chiantini, *On the identifiability of binary segre products*, J. Algebraic Geom. (2011).

[23] L. Brown, *Fundamentals of statistical exponential families: With applications in statistical decision theory*, Institute of Mathematical Statistics, Hayworth, CA, USA, 1986.

[24] C. Carathéodory, *Über den Variabilitätsbereich der Fourierschen Konstanten von positiven harmonischen Funktionen*, Rendiconti del Circolo Matematico di Palermo **32** (1911), 193–217.

[25] M. Á. Carreira-Perpiñan and G. E. Hinton, *On contrastive divergence learning*, Proceedings of the 10-th Interantional Workshop on Artificial Intelligence and Statistics, 2005.

[26] M. V. Catalisano, A. V. Geramita, and A. Gimigliano, *Secant varieties of* $\mathbb{P}^1 \times \cdots \times \mathbb{P}^1$ *(*n*-times) are not defective for* $n \geq 5$, J. Algebraic Geom. **20** (2011), 295–327.

[27] G. Cohen, I. Honkala, S. Litsyn, and A. Lobstein, *Covering codes*, North-Holland mathematical library, Elsevier, 1997.

[28] R. Cordovil and P. Duchet, *Cyclic polytopes and oriented matroids*, European Journal of Combinatorics **21** (2000), no. 1, 49–64.

[29] T. M. Cover and J. A. Thomas, *Elements of information theory, 2nd edition*, John Wiley & Sons, 2006.

[30] H. S. M. Coxeter, *Regular polytopes*, Dover books on advanced mathematics, Dover Publications, 1973.

[31] M. A. Cueto, J. Morton, and B. Sturmfels, *Geometry of the restricted Boltzmann machine*, Algebraic methods in statistics and probability II, AMS Special Session (Marlos A. G. Viana and Henry P. Wynn, eds.), vol. 2, American Mathematical Society, 2010.

[32] M. A. Cueto, E. A. Tobis, and J. Yu, *An implicitization challenge for binary factor analysis*, J. Symb. Comput. **45** (2010), 1296–1315.

[33] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the em algorithm*, Journal of the Royal Statistical Society. Series B (Methodological) **39** (1977), no. 1, 1–38.

[34] M. Develin, *Disjoint faces of complementary dimension*, Contributions to Algebra and Geometry **45** (2004), no. 2, 463–464.

[35] M. Develin and S. Sullivant, *Markov bases of binary graph models*, Annals of Combinatorics **4** (2003), no. 7, 441–466.

[36] P. Diaconis, *Finite forms of de Finetti's theorem on exchangeability*, Synthese **36** (1977), 271–281.

[37] P. Diaconis and D. Freedman, *Finite exchangeable sequences*, The Annals of Probability **8** (1980), 745–764.

[38] P. Diaconis and B. Sturmfels, *Algebraic algorithms for sampling from conditional distributions*, Ann. Statist. **26** (1998), 363–397.

[39] M. P. do Carmo, *Differential geometry of curves and surfaces*, Prentice-Hall, 1976.

[40] A. Dobra, *Markov bases for decomposable graphical models*, Bernoulli **9** (2003), no. 6, 1093–1108.

[41] M. Drton, B. Sturmfels, and S. Sullivant, *Lectures on algebraic statistics*, Oberwolfach Seminars vol. 39, Birkhäuser, 2009.

[42] B. Efron, *The geometry of exponential families*, The Annals of Statistics **6** (1978), no. 2, 362–376.

[43] W. Fenchel, *Über Krümmung und Windung geschlossener Raumkurven*, Math. Ann. **101** (1929), 238–252 (German).

[44] L. Flatto, *A new proof of the transposition theorem*, Proc. American Mathematical Society **24** (1970), no. 1, 29–31.

[45] Y. Freund and D. Haussler, *Unsupervised learning of distributions on binary vectors using 2-layer networks*, NIPS (1992), 912–919.

[46] D. Gale, *Neighborly and cyclic polytopes*, Convexity: Proceedings of the Seventh Symposium in Pure Mathematics of the American Mathematical Society, 1961, pp. 225–233.

[47] E. Gawrilow and M. Joswig, *Polymake: a framework for analyzing convex polytopes*, Polytopes — Combinatorics and Computation (G. Kalai and G. M. Ziegler, eds.), Birkhäuser, 2000, pp. 43–74.

[48] D. Geiger, C. Meek, and B. Sturmfels, *On the toric algebra of graphical models*, Ann. Statist. **34** (2006), 1463–1492.

[49] E. N. Gilbert, *A comparison of signalling alphabets*, Bell System Technical Journal **31** (1952), 504–522.

[50] Z. Gilula, *Singular value decomposition of probability matrices: Probabilistic aspects of latent dichotomous variables*, Biometrika **66** (1979), no. 2, pp. 339–344.

[51] B. Grünbaum, *Convex polytopes*, 2nd ed., John Wiley & Sons, Ltd., 2003.

[52] O. Hanner and H. Rådström, *A generalization of a theorem of Fenchel*, Proceedings of the American Mathematical Society **2** (1951), no. 4, 589–593 (English).

[53] A. Hatcher, *Algebraic topology*, Cambridge University Press, 2002.

[54] M. Henk, J. Richter-Gebert, and G. M. Ziegler, *Basic properties of convex polytopes*, CRC Press, Inc., Boca Raton, FL, USA, 1997.

[55] G. E. Hinton, *Products of experts*, Proceedings of the Ninth Interational Conference on Artificial Neural Networks (ICANN), vol. 1, 1999, pp. 1–6.

[56] G. E. Hinton, *Training products of experts by minimizing contrastive divergence*, Neural Computation **14** (2002), 1771–1800.

[57] ———, *A practical guide to training restricted Boltzmann machines, version 1*, Tech. report, UTML2010-003, University of Toronto, 2010.

[58] G. E. Hinton, S. Osindero, and Y. Teh, *A fast learning algorithm for deep belief nets*, Neural Computation **18** (2006), 1527–1554.

[59] G. E. Hinton and T. J. Sejnowski, *Learning and relearning in boltzmann machines*, pp. 282–317, MIT Press, Cambridge, MA, USA, 1986.

[60] K. J. Horadam, *Hadamard matrices and their applications*, Princeton university press, 2007.

[61] S. Hosten and S. Sullivant, *Gröbner bases and polyhedral geometry of reducible and cyclic models*, J. Combin. Theory Ser. A **100** (2002), 277–301.

[62] J. Håstad and M. Goldmann, *On the power of small-depth threshold circuits*, Foundations of Computer Science, 1990. Proceedings., 31st Annual Symposium on, vol. 2, oct 1990, pp. 610–618.

[63] S. Jukna, *Extremal combinatorics: With applications in computer science*, Texts in Theoretical Computer Science. an EATCS Series, Springer, 2001.

[64] T. Kahle, *Neighborliness of marginal polytopes*, Contributions to Algebra and Geometry **51** (2010), no. 1, 45–56.

[65] T. Kahle and N. Ay, *Support sets of distributions with given interaction structure*, Proceedings of the WUPES, 2006.

[66] T. Kahle, W. Wenzel, and N. Ay, *Hierarchical models, marginal polytopes, and linear codes*, Kybernetika **45** (2009), 189–208.

[67] G. Kalai, *Some aspects of the combinatorial theory of convex polytopes*, 1993.

[68] H. J. Kappen and F. B. Rodríguez, *Efficient learning in Boltzmann Machines using linear response theory*, Neural Computation **10** (1997), 1137–1156.

[69] S. Karlin and W. J. Studden, *Tchebycheff systems: with applications in analysis and statistics*, Pure and applied mathematics, Interscience Publishers, 1966.

[70] T. Krüger, G. F. Montúfar, R. Siegmund-Schultze, and R. Seiler, *Universally typical sets for ergodic sources of multidimensional data*, ArXiv 1105.0393 (2011).

[71] S. Kullback and R.A. Leibler, *On information and sufficiency*, Annals of Mathematical Statistics **22** (1951), 79–86.

[72] N. Le Roux and Y. Bengio, *Representational power of Restricted Boltzmann Machines and Deep Belief Networks*, Neural Computation **20** (2008), no. 6, 1631–1649.

[73] ———, *Deep belief networks are compact universal approximators.*, Neural Computation **22** (2010), 2192–2207.

[74] G. Lebanon, *Axiomatic geometry of conditional models*, IEEE Trans. Inf. Theory **51** (2005), no. 4, 1283–1294.

[75] B. G. Lindsay, *Mixture models: theory, geometry, and applications*, NSF-CBMS regional conference series in probability and statistics, Institute of Mathematical Statistics, 1995.

[76] P. M. Long and R. A. Servedio, *Restricted Boltzmann Machines are hard to approximately evaluate or simulate.*, Proceedings of the 27-th ICML, 2010, pp. 703–710.

[77] B. Sudakov M. Krievelevich and V. H. Vu, *Covering codes with improved density*, IEEE Trans. Inf. Theory **49** (2003), 1812–1815.

[78] B. Matschke, J. Pfeifle, and V. Pilaud, *Prodsimplicial-neighborly polytopes*, Discrete & Computational Geometry (2011), 100–131.

[79] F. Matúš and N. Ay, *On maximization of the information divergence from an exponential family*, Proceedings of the WUPES'03, 2003, pp. 199–204.

[80] G. J. McLachlan and D. Peel, *Finite mixture models*, Wiley series in probability and statistics: Applied probability and statistics, Wiley, 2000.

[81] P. McMullen, *The maximum number of faces of a convex polytope*, Mathematika **XVII** (1970), 179–184.

[82] P. McMullen and G. C. Shephard, *Convex polytopes and the upper bound conjecture*, London Mathematical Society lecture note series, Cambridge University Press, 1971.

[83] R. Memisevic and G. E. Hinton, *Unsupervised learning of image transformations*, In Computer Vision and Pattern Recognition. IEEE Computer Society, 2007.

[84] R. Memisevic and G. E. Hinton, *Learning to Represent Spatial Transformations with Factored Higher-Order Boltzmann Machines*, Neural Computation **22** (2010), no. 6, 1473–1492.

[85] G. F. Montúfar, *Mixture models and representational power of RBMs, DBNs and DBMs*, NIPS Deep Learning and Unsupervised Feature Learning Workshop, 2010.

[86] ———, *Mixture decompositions using a decomposition of the sample space*, Kybernetika (2011), accepted.

[87] G. F. Montúfar and N. Ay, *Refinements of universal approximation results for Deep Belief Networks and Restricted Boltzmann Machines*, Neural Computation **23** (2011), no. 5, 1306–1319.

[88] G. F. Montúfar, J. Rauh, and N. Ay, *Expressive power and approximation errors of Restricted Boltzmann Machines*, Advances in Neural Information Processing Systems 24 (J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F.C.N. Pereira, and K.Q. Weinberger, eds.), MIT Press, 2011, pp. 415–423.

[89] P. C. Ojha, *Enumeration of linear threshold functions from the lattice of hyperplane intersections*, Neural Networks, IEEE Transactions on **11** (2000), no. 4, 839 –850.

[90] R. Osserman, *A survey of minimal surfaces*, Dover books on advanced mathematics, Dover Publications, 1986.

[91] P. R. J. Östergård and M. K. Kaikkonen, *New upper bounds for binary covering codes*, Discrete Mathematics **178** (1998), no. 1-3, 165 – 179.

[92] K. Pearson, *Contributions to the mathematical theory of evolution*, Philosophical Transactions of the Royal Society A **185** (1894), 71–110.

[93] G. Pólya and G. Szegő, *Aufgaben und Lehrsätze aus der Analysis*, Heidelberger Taschenbücher, no. II, Springer, 1970 (German).

[94] K. Ranestad and B. Sturmfels, *On the convex hull of a space curve*, Advances in Geometry (2011).

[95] J. Rauh, *Finding the maximizers of the information divergence from an exponential family*, Ph.D. thesis, Universität Leipzig, 2011.

[96] J. Rauh, T. Kahle, and N. Ay, *Support sets of exponential families and oriented matroids*, International Journal of Approximate Reasoning **52** (2011), no. 5, 613–626.

[97] S. Ray and B. G. Lindsay, *The topography of multivariate normal mixtures*, The Annals of Statistics **33** (2005), no. 5, 2042–2065.

[98] W. E. Roth, *On direct product matrices*, Bulletin of the American Mathematical Society **40** (1934), 461–468.

[99] R. Salakhutdinov and G. E. Hinton, *Deep Boltzmann Machines*, Artificial Intelligence **5** (2009), no. 2, 448455.

[100] B. Sallans and G. E. Hinton, *Reinforcement learning with factored states and actions*, J. Mach. Learn. Res. **5** (2004), 1063–1088.

[101] A. Sard, *The measure of the critical values of differentiable maps*, Bulletin of the American Mathematical Society **48** (1942), 883–890.

[102] T. J. Sejnowski, *Higher-order boltzmann machines*, Neural Networks for Computing, American Institute of Physics, 1986, pp. 398–403.

[103] R. Settimi and J. Q. Smith, *On the geometry of bayesian graphical models with hidden variables*, Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence, UAI'98, Morgan Kaufmann Publishers Inc., 1998, pp. 472–479.

[104] M. Shaked, *On mixtures from exponential families*, Journal of the Royal Statistical Society. Series B (Methodological) **42** (1980), no. 2, 192–198.

[105] I. Shemer, *Neighborly polytopes*, Israel Journal of Mathematics **43** (1982), 291–311.

[106] R. C. Singleton, *Maximum distance q-nary codes*, IEEE Trans. Inf. Theory **10** (1964), no. 2, 116–118.

[107] P. Smolensky, *Information processing in dynamical systems: foundations of harmony theory*, Symposium on Parallel and Distributed Processing, 1986.

[108] B. Sturmfels, *Cyclic polytopes and d-order curves*, Geometriae Dedicata **24** (1987), 103–107.

[109] ———, *Gröbner bases and convex polytopes*, University lecture series, American Mathematical Society, 1996.

[110] I. Sutskever and G. E. Hinton, *Deep narrow sigmoid belief networks are universal approximators.*, Neural Computation **20** (2008), 2629–2636.

[111] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning)*, The MIT Press, March 1998.

[112] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, *Policy gradient methods for reinforcement learning with function approximation*, Adv. in NIPS **12** (2000), 1057–10063.

[113] T. Tanaka, *Information geometry of mean-field approximation*, Neural Computation **12** (2000), 1951–1968.

[114] D. M. Titterington, A. F. M. Smith, and U. E. Makov, *Statistical analysis of finite mixture distributions*, John Wiley and Sons, 1985.

[115] P. E. Utgoff and D. J. Stracuzzi, *Many-layered learning*, Neural Computation **14** (2002), 2002.

[116] R. R. Varshamov, *Estimate of the number of signals in error correcting codes*, Doklady Akad. Nauk SSSR **117** (1957), 739–741.

[117] N. N. Čencov, *Statistical decision rules and optimal inference*, vol. 53, American Mathematical Society, 1972, Translations of mathematical monographs.

[118] M. J. Wainwright and M. I. Jordan, *Graphical models, exponential families, and variational inference*, vol. 1, now Publishers Inc. Foundations and Trends in Machine Learning, 2008.

[119] S. Watanabe, *Algebraic geometry and statistical learning theory*, Cambridge monographs on applied and computational mathematics, Cambridge University Press, 2009.

[120] S. W. Weis, *Exponential families with incompatible statistics and their entropy distance*, Ph.D. thesis, Friedrich-Alexander-Universität Erlangen-Nürnberg, 2009.

[121] K. G. Zahedi, N. Ay, and R. Der, *Higher coordination with less control – a result of infromation maximization in the sensomotor loop*, Adaptive Behavior **18** (2010), no. 3-4, 338–355.

[122] G. M. Ziegler, *Lectures on polytopes*, Graduate texts in mathematics, Springer-Verlag, 1995.

[123] P. Zwiernik and J. Q. Smith, *Implicit inequality constraints in a binary tree model*, Electronic Journal of Statistics **5** (2011), 1276–1312.

Index

List of Symbols

$[n]$	The set $\{1, \dots, n\}$ for some $n \in \mathbb{N}$
$[y_\lambda]$	The cylinder set $\{x \in \times_{i\in[n]} \mathcal{X}_i \colon x_i = y_i \ \forall i \in \lambda\}$ for some $\lambda \subseteq [n]$
$\mathbb{1}$	Constant function $\mathbb{1} \equiv 1$
$\mathbb{1}_{\mathcal{Y}}$	Characteristic function on $\mathcal{Y}$
$2^{[n]}$	The power set of $[n]$
A	A sufficient statistics matrix with columns $A_x \in \mathbb{R}^d$ for $x \in \mathcal{X}$ and rows $A_i \in \mathbb{R}^{\mathcal{X}}$, page 14
$\mathcal{A}_q(n, d)$	Maximal cardinality of a q-ary code of length n and minimum distance d, page 21
$\mathrm{Car}_V(V')$	Carathéodory number of V' w.r.t. V, page 65
C_n	Poset of face-incident vertices of the n-dimensional unit cube
$\mathrm{conv}(\mathcal{M})$	Convex hull of the set $\mathcal{M}$
$\mathrm{cs}(\mathcal{E})$	Convex support (marginal polytope) of the exponential family $\mathcal{E}$
$C(v, d)$	Cyclic polytope with v vertices and dimension d, page 17
$\mathrm{DBN}_{n_0,n_1,\dots,n_l}$	Deep Belief Network model with n_0 visible units and l hidden layers of widths $n_1, \dots, n_l$, page 74
$d_H(\cdot, \cdot)$	Hamming distance
$\mathcal{E}(\mathcal{X})$	Exponential family supported by $\mathcal{X}$, page 14
$\partial\mathcal{E}$	The boundary of the exponential family $\mathcal{E}$, page 14
$\mathcal{E}^k(\bigtimes_{i=1}^n \mathcal{X}_i)$	k-Interaction exponential family on $\mathcal{X} = \bigtimes_{i=1}^n \mathcal{X}_i$, page 15
$\mathcal{E}^k_{n,q\text{-ary}}$	k-Interaction exponential family on $\mathcal{X} = \mathbb{F}_q^n$
η	Expectation parameter
$\mathrm{ex}(\mathcal{M})$	The extreme points of the set $\mathcal{M}$
$\mathcal{F}(Q)$	Face lattice of a polytope Q, page 16
$\mathcal{F}(\mathcal{E})$	Facial sets of the exponential family $\mathcal{E}$, page 17
k-Hamiltonian	Denotes a model which contains $\mathrm{skel}_{k-1}(\overline{\mathcal{P}})$, page 57

$\kappa^f_{\mathcal{E}}$	Minimal cardinality of a facial packing, page 18
$\kappa^s_{\mathcal{E}}$	Minimal cardinality of an S-set covering, page 18
$K(n, R)$	The smallest cardinality of an n-bit code with covering radius R, page 27
$\mathrm{Mixt}(\mathcal{M}_1, \ldots, \mathcal{M}_m)$	Mixture with basis points in $\mathcal{M}_1$ through $\mathcal{M}_m$
$\mathrm{Mixt}^m(\mathcal{M})$	m-th mixture of $\mathcal{M}$, page 13
$\mathcal{N}$	Normal space of an exponential family
ν	A reference measure, page 14
$\overline{\mathcal{P}}(\mathcal{X})$	Set of all probability distributions on $\mathcal{X}$
$\mathcal{P}(\mathcal{X})$	Set of strictly positive probability distributions on $\mathcal{X}$
$\pm$	Indicates that the statement is true inserting $+$ or $-$ in the place of $\pm$
$\mathcal{P}_n \equiv \mathcal{P}_{n,\mathrm{bin}} \equiv \mathcal{P}(\{0,1\}^n)$	
$\mathcal{P}_\xi$	Partition model with partition ξ, page 90
$\mathrm{RBM}_{n,m}$	Restricted Boltzmann Machine model with n visible and m hidden units, page 73
$\mathrm{ri}(\mathcal{M})$	The relative interior of the set $\mathcal{M}$
$\mathbb{R}^{\mathcal{X}}$	The space of real-valued functions on $\mathcal{X}$
$\mathbb{R}_>$	The reals strictly larger than 0
$\mathcal{S}_{\nu,\varrho}$	Convex exponential family with reference measure ν and partition ϱ, page 50
S-set	A set $\mathcal{Y} \subseteq \mathcal{X}$ s.t. for a given model $\mathcal{M}$, $\mathcal{M} \supseteq \overline{\mathcal{P}}(\mathcal{Y})$, page 18
u	The uniform probability distribution
$u_{\mathcal{Y}}$	The uniform probability distribution on the set $\mathcal{Y}$
$\dot\cup$	Disjoint union
$\mathcal{X}$	A finite set
$Z_{+,n}$	Set of length-n binary vectors with an even number of ones, page 9

Acknowledgments

I am deeply grateful to Nihat Ay for the supervision of this thesis, steady guidance, trust, and support during my work and undertakings of the past years.
I am grateful to Johannes Rauh, who has been a great coworker and skilled discussion partner for all sorts of ideas.
I am grateful to Jürgen Jost for support and advice, and the the work climate at MPI MIS, which greatly benefited my research.
I am deeply grateful to Shun-ichi Amari for stimulating and encouraging discussions, and for sharing his vast range of ideas during my time at RIKEN BSI.
I am deeply grateful to Yoshua Bengio for his interest in my work and for instructive discussions about Deep Learning.
I am grateful to Bernd Sturmfels for stimulating discussions and for using his talent to connect people and mathematical problems.
I thank Jason Morton for his interest in my work and for stimulating discussions.
I am grateful to my colleges at MPI MIS, Areejit Samal, Alihan Kabalak, Stephan Poppe, Pierre-Yves Bourguignon, Keyan Zahedi, Camilo Sarmiento, Chiranjib Mukerjee, Stephan Weis, and jj-all, who have been great companions. I thank Thomas Kahle for valuable discussions on combinatorics of polytopes and exponential families.
I am over-grateful to Antje Vandenberg for steady assistance and for making bureaucratic procedures so simple.
I thank Heike Rackwitz, the library group, the computer group, and the administration at MPI MIS.
I am grateful to the International Max Planck Research School Mathematics in the Sciences and the Research Academy Leipzig.
A whole-hearted thanks goes to my family, friends and to Katherina.

Bibliographische Daten

On the Expressive Power of Discrete Mixture Models, Restricted Boltzmann Machines, and Deep Belief Networks—A Unified Mathematical Treatment

Übersetzung ins Deutsche: Über die Darstellungskraft diskreter Mischungsmodelle, eingeschränkter Boltzmann-Maschinen und tiefer Bayes'scher Netzwerke – Eine vereinheitlichte mathematische Behandlung –

Montúfar Cuartas, Guido Francisco
Universität Leipzig, Dissertation, 2012

155 Seiten, 30 Abbildungen, 123 Referenzen

Information about the author

Name: Guido Francisco Montúfar Cuartas
Date of Birth: May 13, 1983 in Panama City, Panama
Education: 2009–2012 Ph.D. studies in mathematics. Universität Leipzig, Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany
Dec. 2008 Diplom Physiker
2004–2008 Physics studies. Technische Universität Berlin, Germany.
Aug. 2007 Diplom Mathematiker
2002–2007 Mathematics studies. Technische Universität Berlin, Germany.