Max-Planck-Institut
für Mathematik
in den Naturwissenschaften
Leipzig

Mixture Decomposition of Distributions using a Decomposition of the Sample Space

by

Guido Montúfar

Preprint no.: 39 2010

Mixture Decomposition of Distributions using a Decomposition of the Sample Space

Guido Montufar*
Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany.

Abstract

We consider the set of join probability distributions of N binary random variables which can be written as a sum of m distributions in the following form $p(x_1, \ldots, x_N) = \sum_{i=1}^{m} \alpha_i f_i(x_1, \ldots, x_N)$, where $\alpha_i \geq 0$, $\sum_{i=1}^{m} \alpha_i = 1$, and the $f_i(x_1, \ldots, x_N)$ belong to some exponential family. For our analysis we decompose the sample space into portions on which the mixture components f_i can be chosen arbitrarily. We derive lower bounds on the number of mixture components from a given exponential family necessary to represent distributions with arbitrary correlations up to a certain order or to represent any distribution. For instance, in the case where f_i are independent distributions we show that every distribution p on $\{0,1\}^N$ is contained in the mixture model whenever $m \geq 2^{N-1}$, and furthermore, that there are distributions which are not contained in the mixture model whenever $m < 2^{N-1}$.

1 Introduction

A probability mixture model is a set of distributions which can be written as convex combination of other distributions belonging to a family of distributions. The idea is that the sum of parts which are individually relatively easy to describe can result in a powerful and versatile machinery. Mixture models have a long history, and there has given many advances in their study, e.g. the identifiablity and mixture density estimation problems have been tackled with the familiar method of moments, and the expectation maximization algorithm. Mixture models have also found a wide range of applications, e.g. in clustering and machine learning and many others, see for instance D. M. Titterington et al. (1985), B. G. Lindsay (1995). However, important questions, particularly about the dimension of mixture models or their representational power still remain open, M. Drton et al. (2009). In this paper we focus on the representational power of mixture models, i.e., we ask how large must a mixture of simple distributions be in order to contain families of more complicated correlated distributions. As an example of this kind of problems one may think of the decomposition of exchangeable

* montufar@mis.mpg.de

distributions as convex combinations of Bernoulli distributions, P. Diaconis (1977). Rather than using a decomposition in extremal points (distributions which can only be trivially decomposed) we use elements from convex sections of the boundary of the family the mixture components belong to. There are such boundary sections of exponential families, which can be identified with subsets of the sample space, as we will explain below. This idea builds on the previous works J. Rauh et al. (2009); T. Kahle (2010, 2006); Geiger et al. (2006).

In discrete mixture models a family of distributions $\mathcal{E} \subseteq \overline{\mathcal{P}(\mathcal{X})}$ is given, where $\overline{\mathcal{P}(\mathcal{X})}$ is the set of all join distributions of N random variables $(X_1, \ldots, X_N) =: X$ with sample space $\mathcal{X} = \times_{i=1}^{N}[r_i]$ for some natural numbers r_i, $[r_i]$ beeing a set containing r_i elements. For straightforwardness we consider binary variables, i.e. $\mathcal{X} = \{0,1\}^N$. A natural way to understand mixture models, M. Drton et al. (2009), is to assume that there is a hidden random variable Y with state space $[m]$, and that for each $y \in [m]$, a mixture component is given by the conditional distribution of X given $Y = y$, $p_y \in \mathcal{E}$. If the random variable Y has distribution $\alpha \in \overline{\mathcal{P}([m])}$, then the join distribution of Y and X is given by

$$\Pr(Y = y, X = x) = \alpha(y)\, p_y(x).$$

Since the variable Y is assumed to be hidden, only the marginal distribution of X is visible, i.e.

$$\Pr(X = x) = \sum_{y=1}^{m} \alpha(y)\, p_y(x).$$

Suppose for example that the mixture components can be chosen arbitrarily from $\mathcal{E} = \{\delta_y\}_y$. Then, the convex combinations of the form

$$\sum_y \alpha(y)\, \delta_y(x)$$

cover all distributions in $\overline{\mathcal{P}}$ if there are as many y as x. This is simply a direct parametrization of a distribution in terms of its values on the different x. On the other hand, this model has $2^N - 1 = |\mathcal{X}| - 1$ parameters and it is clear that a smaller number of mixture components would not suffice to represent some distributions. More generally, a problem arises when α cannot be chosen arbitrarily, but in some further model. This can be the case when it comes to approximate probability distributions as marginals say of Restricted Boltzmann Machines. We will comment on this at the end of this section. In this paper we focus on the simpler case where the mixture weights α can be chosen arbitrarily and ask what happens when one allows more general mixture components than $\{\delta_x\}$. How many mixture components from a certain model are required / sufficient if we want to represent any distribution, or respectively, distributions from the class describing correlations of a certain order?

We consider here mixtures with components from the set of independent distributions of N binary random variables, called the independence model, and which consists of all factorizable distributions. And, more generally, we consider mixtures from some exponential family. Very familiar examples of exponential families are multivariate gaussians or just the independent distributions. There is a hierarchy of exponential families

$$\mathcal{E}^1 \subset \mathcal{E}^2 \subset \cdots \subset \mathcal{E}^N = \mathcal{P}(\mathcal{X}),$$

where for $\mathcal{E}^k$ the index k denotes the order of correlations between the N random variables that are covered by $\mathcal{E}^k$, see for instance S. Amari (1999). We will explain these objects in more detail in the next section.

The mixture models that we study are of the following form:

$$\mathcal{M}_m^k := \left\{ \sum_{j=1}^{m} \alpha_j f_j \in \mathbf{R}^{|\mathcal{X}|} : \alpha_i \geq 1, \sum \alpha_i = 1 \text{ and } f_i \in \overline{\mathcal{E}^k} \text{ for all } j \right\} \subseteq \overline{\mathcal{P}},$$

where $\overline{\mathcal{E}^k}$ is the closure of $\mathcal{E}^k$ in $\mathbf{R}^{|\mathcal{X}|}$.

For any $1 \leq k \leq N$ the set $\overline{\mathcal{E}^k}$ contains the atoms $\{\delta_x\}_{x\in\mathcal{X}}$, and since these are the extremal points of $\overline{\mathcal{P}}$, any distribution can be represented as a mixture of $|\mathcal{X}|$ elements of $\overline{\mathcal{E}^k}$, (when the mixture weights can be chosen arbitrarily), as explained above. In our notation we can state:

$$\mathcal{M}_{m\geq|\mathcal{X}|}^k = \overline{\mathcal{P}(\mathcal{X})}, \quad \forall 1 \leq k \leq N.$$

Now, the first question is whether a smaller number of mixture components m suffices, depending on k. The second question is how small can m be when it is only required that

$$\mathcal{M}_m^k \supseteq \overline{\mathcal{E}^l}.$$

Can we derive relations between l, k and m? How do these relations depend on the number of random variables N? The subject of this note is to derive such relations. Among many others this questions are of high relevance for understanding stochastic networks like Restricted Boltzmann Machines or Deep Belief Networks, see for instance Montufar & Ay (2010), and for the description of correlated neural spikes, see for instance S. Amari (2010).

The central idea of this paper is to find decompositions of the sample space such that all distributions with support contained in the members of this decomposition are contained in the set of distributions where the mixture components are taken from. This allows a very simple decomposition of distributions, and furthermore, when these support sets are given and chosen disjointly, then the identification problem is automatically solved (uniqueness of the mixture representation within the model).

To motivate our ansatz we want to review a small example: The mixtures of two independent distributions of two binary variables. The set of mixtures of two fixed elements from the independence model can be represented as a line connecting the two elements. If the two elements are not fixed, the set of lines connecting points on the independence model for $\mathcal{X} = \{0, 1\}^2$ covers all of the probability simplex, (see Figure 1 in the next section). The situation becomes more complicated for N larger than 2, since the dimension of the independence model increases only as the logarithm of the dimension of the probability simplex. However, a closer inspection reveals that mixtures of two elements lying in the intervals $[\delta_{(0,1)}, \delta_{(1,1)}]$ and $[\delta_{(1,0)}, \delta_{(0,0)}]$, which in fact belong to $\overline{\mathcal{E}^1}$, already suffice to cover all the probability simplex. The sets of distributions described by these intervals have the special property that they comprend all possible distributions with support sets $\{(0,1),(1,1)\}$ and $\{(1,0),(0,0)\}$ and that the union of those two sets is $\mathcal{X}$.

In this paper we elaborate that observation for the general setting. This reasonings allow us to show for example that all distributions with support restricted to some special sets are contained in the independence model and that $2^N/2$ such sets cover all the state space $\{0,1\}^N$ for arbitrary N. This can be directly used to decompose arbitrary distributions as mixtures of independent distributions.

The same questions posed above can be considered for the case when the mixture weights belong to a certain model. Setting

$$^{n}\mathcal{M}_m^k := \left\{ p(x) = \sum_{j=1}^{m} \alpha_j f_j(x) : f_j \in \overline{\mathcal{E}^k}\,\forall j, \alpha = (\alpha_1, \ldots, \alpha_m) \in \overline{\mathcal{E}^n([m])} \right\},$$

can we find sufficient / necessary conditions on n, l, m, k such that

$$^{n}\mathcal{M}_m^k \supseteq \overline{\mathcal{E}^l}\,?$$

Montufar & Ay (2010) showed that in the case of mixture weights from the independence model, $\alpha \in \overline{\mathcal{E}^1}$, a number $2^{2^{N-1}}$ of mixture components from the independence model suffices to represent any distribution on $\{0,1\}^N$, i.e.,

$$^{1}\mathcal{M}^1_{m=2^{2^{N-1}}} = \overline{\mathcal{P}(\{0,1\}^N)}.$$

This number of mixture components is very large, but it is far from trivial to prove or disprove the optimality of this result. We hope that the ideas developed in this paper will also help to approach this kind of problems.

2 Preparations

In this section we present concepts and results needed for the proofs of our main results in the next section.

Given a family of sets $\Delta \subseteq 2^{[N]}$, $2^{[N]}$ the power set of $[N] := \{1, \ldots, N\}$, and a matrix $(A_{\lambda,x})_{\lambda\in\Delta, x\in\mathcal{X}}$, we define a model $\mathcal{E}_\Delta := \{p(x) = \exp\left(\sum_{\lambda\in\Delta} J_\lambda A_{\lambda,x}\right) : J \in \mathbf{R}^\Delta\}$, which in the literature is known as an *exponential family*. Here we assume that $\emptyset \in \Delta$ and $A_{\emptyset,x} = 1$ for all x, such that $J_\emptyset$ is a constrained parameter which normalizes the distribution. For $\mathcal{X} = \{0,1\}^N$ the model $\overline{\mathcal{E}_\Delta}$ is a subset of the $2^N - 1$-dimensional simplex (the set of all distributions on binary vectors of length N) in $\mathbf{R}^{2^N}$. The closure of this set is denoted by $\overline{\mathcal{E}_\Delta}$. This closure contains in particular distributions which do not have full support.

An exponential family defined this way is completely characterized by the row-span of the matrix $(A_{\lambda,x}) = [A_{x^1}, \ldots, A_{x^{|\mathcal{X}|}}]$, where $A_x = (A_{\lambda_1,x}, \ldots, A_{\lambda_{|\Delta|},x})^t$. An *independence model* or *set of independent distributions* is described by the exponential family for which $\Delta = [N] \cup \emptyset$, and $A_{\lambda,x} = x_\lambda$ for $\lambda \neq \emptyset$ and $A_{\emptyset,x} = 1$. In this case we have namely that $p(x_1, \ldots, x_N) = \exp\left(J_\emptyset + \sum_{i=1}^N J_i x_i\right) = \exp(J_\emptyset) \prod_{i=1}^N \exp(J_i x_i)$, with arbitrary values J_i. This model, as a subset of $\mathbf{R}^{2^N}$ is a manifold of dimension N, with canonical coordinate functions $\{J_i\}_{i\in[N]}$. For simplicity we write $\mathcal{E}^1$ for the independence model. A distribution in the boundary of $\mathcal{E}^1$ is for example the following: $p(x_1, x_2) = p^1(x_1) \cdot p^2(x_2)$, where p^1 assigns probability one to $\{x_1 = 1\}$, and p^2 assigns probability ω to $\{x_2 = 1\}$. This distribution vanishes in $(x_1, x_2) = (0, 0)$.

For the description of correlated distributions we consider first the space of functions of N variables $\{x_i\}_{i=1}^N$ describing arbitrary interactions of any subset $\{x_i\}_{i\in\Lambda}$ with indices $\Lambda \in \Delta \subseteq 2^N$, see for instance T. Kahle (2006):

$$I_\Delta := \left\{ g \in \mathbf{R}^{\mathcal{X}} : g = \sum_{\Lambda\in\Delta} g_\Lambda, \text{ where } g_\Lambda \text{ is a function of } \{x_i\}_{i\in\Lambda} \right\}.$$

The exponential family describing distributions with correlations specified by Δ is then the set of distributions of the form $p \propto \exp(g)$, $g \in I_\Delta$. Notice that the functions $\{\sigma_\lambda(x_1, \ldots, x_N) := (-1)^{|\{i\in\lambda : x_i = 1\}|}\}_{\lambda\subseteq\Lambda}$, (known as characters), form a basis of the functions of the binary variables $\{x_i\}_{i\in\Lambda}$ for any $\Lambda \subseteq [N]$. Since an exponential family as described above is characterized by the row-span of A, the distributions with arbitrary correlations up to order k are given by the following family:

$$\mathcal{E}^k := \left\{ p(x) = \exp\left(\sum_{\lambda\in\Delta_k} J_\lambda A_{\lambda,x} \right), J \in \mathbf{R}^{|\Delta_k|} \right\},$$

where $A_{\lambda,x} = (-1)^{|\{i\in\lambda : x_i = 1\}|}$, $\Delta_k := \{\lambda \in 2^N : |\lambda| \leq k\}$, and $J_\emptyset$ normalizes the distributions.

The *marginal polytope* of the model $\overline{\mathcal{E}_\Delta}$ is $Q_\Delta := \mathrm{conv}\{A_x\}_{x\in\mathcal{X}}$. This is the convex hull of the column vectors A_x in $\mathbf{R}^{|\Delta|}$. For simplicity we write Q_k for Q_{Δ_k}. A *face* of Q_Δ is the intersection of Q_Δ with an hyperplane of codimension one in $\mathbf{R}^{|\Delta|}$ such that all points of Q_Δ lie on one of the closed halfspaces defined through that hyperplane. The marginal polytope contains information about the support-sets of the distributions contained in the model, as will be explained below. See Figure 1.

Definition 1. (Facial sets) *Given the model $\overline{\mathcal{E}_\Delta}$, a Q_Δ-*facial *set is a set $\mathcal{Y} \subseteq \mathcal{X}$ such that* $\mathrm{conv}\{A_x\}_{x\in\mathcal{Y}}$ *is a face of Q_Δ.*

From Figure 1 we gather that the facial sets in the independence model for $\mathcal{X} = \{0, 1\}^2$ are $\{(1, 1), (1, 0), (0, 1), (0, 0)\}$, as well as $\{(1, 1), (0, 1)\}$, $\{(1, 1), (1, 0)\}$ and $\{(0, 0), (0, 1)\}$, $\{(0, 0), (1, 0)\}$, and all sets of cardinality one.

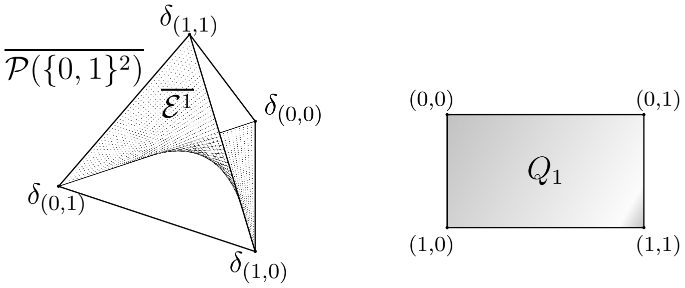

Figure 1: On the left, the probability simplex for the sample space $\mathcal{X} = \{0,1\}^2$, and the independence model. On the right, the marginal polytope of the independence model (omitting the first coordinate, which is allways 1).

Proposition 2. (Facial Sets, J. Rauh et al. (2009)) *If $\mathcal{Y} \subseteq \mathcal{X}$ is Q_Δ-facial, then there exists one $p \in \overline{\mathcal{E}_\Delta}$ with $\operatorname{supp}(p) = \mathcal{Y}$. Furthermore, if $p \in \overline{\mathcal{E}_\Delta}$ then $\operatorname{supp}(p)$ is facial.*

This proposition tells us in particular that a distribution in the model $\overline{\mathcal{E}_\Delta}$ can only have support $\mathcal{Y}$, if $\mathcal{Y}$ is Q_Δ- facial. This yields the following:

Proposition 3. (Support sets in the Independence Model) *In the independence model $\overline{\mathcal{E}^1}$,we have $Q_1 = \operatorname{conv}\{\hat{x}\}_{x\in\mathcal{X}}$, where $\hat{x} = (1, x_1, x_2, \dots, x_N)^t$, which corresponds to the N-cube. The only sets which can occurr as support of distributions in the closure of the independence model consist of the sets of binary vectors of length N whose convex hull is a face of the N-cube.*

Proposition 2 also tells us that whenever a set $\mathcal{Y}$ is facial, there exists a distribution in the model which has $\mathcal{Y}$ as its support. This does not mean that all distributions with support $\mathcal{Y}$ are contained in the model. This motivates the following definition:

Definition 4. (S-sets) *Given a model, i.e. a set of probability distributions on $\mathcal{X}$, we say that a set $\mathcal{Y} \subseteq \mathcal{X}$ has the S-*property *or is an S-*set *in that model if and only if every distribution with support $\mathcal{Y}$ is contained in the model.*

For the independence model we have the following:

Proposition 5. (S-sets in the Independence Model) *A set has the S-property in the closure of the independence model if and only if it has cardinality one or consists of two binary vectors which differ in exactly one entry.*

The statement is a special case of Lemma 7, a characterization of S-sets as the facial sets for which the corresponding face of the marginal polytope is a simplex, in view of the fact that the marginal polytope of the independence model is the N-cube, Proposition 3. Here we present an alternative proof of the *if* part of Proposition 5 which does not require concepts of marginal polytopes (similar to the proof of Theorem 1, part 3 in Montufar & Ay (2010)).

Proof of Lemma 5, 'if' part. Let e_j be the vector with value one in the position j and zeros elsewhere, and let $\mathbb{1}$ be the vector consisting of ones everywhere. For any binary vector v we we write $v_{\hat{j}}$ for the binary vector which is equal to v everywhere but in the entry j is different. Regarad that any element of the independence model is of the form $f(x) \propto \exp(w \cdot x + c)$, where c is arbitrary. Now, for some arbitrary $j \in [N]$ let $\tilde{x}$ be an arbitrary vector in $\mathcal{X}$ with $\tilde{x}_j = 1$. Consider some $a \in \mathbf{R}$, and any $\lambda_1, \lambda_2 \in \mathbf{R}$. Define $\hat{w} := a(\tilde{x}_{\hat{j}} - \frac{1}{2}\mathbb{1}_{\hat{j}})$, $\bar{w} := \hat{w} + (\lambda_2 - \lambda_1)e_j$, and $\bar{c} := -\hat{w} \cdot \tilde{x} + \lambda_1 = -\hat{w} \cdot \tilde{x}_{\hat{j}} + \lambda_1$. Set $s = |\mathrm{supp}\tilde{x}_{\hat{j}}|$ (number of entries of $\tilde{x}_{\hat{j}}$ with value one). For the parameters $\bar{w}$ and $\bar{c}$ we have:

$$\begin{aligned}\bar{w} \cdot x &= \frac{1}{2}a(s - |\{i : (\tilde{x}_{\hat{j}})_i \neq (x_{\hat{j}})_i\}|) + (\lambda_2 - \lambda_1)x_j, \\ \bar{c} &= -\frac{1}{2}as + \lambda_1.\end{aligned}$$

In the limit $a \to \infty$ we get $\exp(\bar{w} \cdot x + \bar{c}) = 0 \ \forall x \neq \tilde{x}, \tilde{x}_{\hat{j}}$, and $\exp(\bar{w} \cdot \tilde{x}_{\hat{j}} + \bar{c}) = e^{\lambda_1}$, and $\exp(\bar{w} \cdot \tilde{x} + \bar{c}) = e^{\lambda_2}$. This is $f(x) = \lim_{a \to \infty} \exp(\bar{w}x + \bar{c})/Z$, ($Z$ the normalization), vanishes everywhere but in an arbitrary pair of vectors which differ in exactly one entry. In this pair of vectors f takes arbitrary values (adding to one). □

We will use the following result established by T. Kahle (2010). Given an interaction set $\Delta \subseteq 2^{[N]}$, a smallest set $\lambda \in 2^{[N]}$ not belonging to Δ is called a minimal non-face of Δ.

Theorem 6. (Neighborliness of marginal polytopes, Kahle 2010) *Let $k + 1$ be the minimal cardinality among the non-faces of Δ, then every probability distribution p with $|\operatorname{supp} p| < 2^k$ is contained in $\overline{\mathcal{E}_\Delta}$.*

This theorem states that all sets of cardinality smaller than 2^k are S-sets in the model $\overline{\mathcal{E}^k}$. This result is optimal in the sense that there exists distributions with support of cardinality 2^k, which are not contained in $\overline{\mathcal{E}^k}$. Theorem 6 can be expressed in terms of the marginal polytope Q_k, by saying that it is $(2^k - 1)$-neighborly. This means that any set vertices of Q_k containing at most $(2^k - 1)$ elements are the extremal points of a face of Q_k. This implies that any $(2^k - 1)$ vertices describe a face of the marginal polytope which is a simplex. This obsevation motivates a characterization of S-sets in terms of faces of the marginal polytope which are simplices, what we do in Lemma 7 in the next section.

3 Mixture Decompositions

The following two results represent the reasonings of this paper.

Lemma 7. (Characterization of the S-property) *A set $\mathcal{Y} \subseteq \mathcal{X}$ has the S-property in the exponential family defined by the matrix A, and that model contains all distributions with support contained in $\mathcal{Y}$, if and only if $\operatorname{conv}\{A_x\}_{x \in \mathcal{Y}'}$ is a face of $\operatorname{conv}\{A_x\}_{x \in \mathcal{X}}$ for all $\mathcal{Y}' \subseteq \mathcal{Y}$.*

The lemma above says that every distribution with support $\mathcal{Y}$ is contained in the exponential family described by A exactly when the vertices of the marginal polytope corresponding to $\mathcal{Y}$ are the extremal points of a face of the marginal polytope which is a simplex. The proof of Lemma 7 is in the Appendix.

The idea of the following lemma is to decompose an arbitrary distribution as a mixture of distributions with support set given by S-sets (which by the lemma above are directly related to special faces of the marginal polytope) and such that the union of their support sets is $\mathcal{X}$.

Lemma 8. (Mixture decompositions using a decomposition of the sample space) *Given any $p \in \overline{\mathcal{P}(\mathcal{X})}$, there exist $f_i \in \overline{\mathcal{E}}$, and $\alpha_i \geq 0$, $\sum_i \alpha_i = 1$, $i = 1, \ldots, \kappa$ such that*

$$p(x) = \sum_{i=1}^{\kappa} \alpha_i f_i(x),$$

where $\kappa = \min\{|\{\mathcal{Y}_i\}_i| : \forall i\, \mathcal{Y}_i$ is an S-set in $\overline{\mathcal{E}},$ and $\cup_i \mathcal{Y}_i = \mathcal{X}\}$.

Since any $f_i \in \overline{\mathcal{E}}$ is arbitrarily well approximated by an element in $\mathcal{E}$, any $p \in \overline{\mathcal{P}}$ is arbitrarily well apprixmated as mixture of κ elements in $\mathcal{E}$.

Proof of Lemma 8. The f_i can be chosen arbitrarily with the only restriction that its support is confined to an S-set $\mathcal{Y}_i$. We can make the $\mathcal{Y}_i$ disjoint while $\cup_i \mathcal{Y}_i = \mathcal{X}$, since any subset of an S-set is again an S-set. Setting $f_i(x) = p(x)/p(\mathcal{Y}_i)$ for $x \in \mathcal{Y}_i$, and $f_i(x) = 0$ for $x \notin \mathcal{Y}_i$, and $\alpha_i = p(\mathcal{Y}_i)$ yields the result. □

Mixtures from the independence model

Proposition 5 provides an easy way to decompose distributions as mixtures of elements in the independence model:

Theorem 9. (Mixtures of independent distributions) *Given any distribution p on the binary vectors of length N, there exist 2^{N-1} elements in the independence model $f_i \in \overline{\mathcal{E}^1}$, and weights $\alpha_i \geq 0$, $\sum_i \alpha_i = 1$ such that*

$$p(x) = \sum_{i=1}^{2^N/2} \alpha_i f_i(x).$$

This also implies that a mixture of 2^{N-1} elements from $\mathcal{E}^1$ approximates any distribution in $\overline{\mathcal{P}}$ arbitrarily well. The special case of our Theorem 9 where $N = 2$ is the content of Theorem 2 by S. Amari (2010).

Proof of Theorem 9. Select a perfect matching of the graph of the N-cube $\{x^{i,1}, x^{i,2}\}_i$. This consists of $\frac{2^N}{2}$ disjoint pairs of points (edges) covering all vertices of the N-cube. From Lemma 5 we have that to every i, all distributions with support $\{x^{i,1}, x^{i,2}\}$ are contained in $\overline{\mathcal{E}^1}$. Choose now $\alpha_i = p(x^{i,1}) + p(x^{i,2})$, and f_i with support $\{x^{i,1}, x^{i,2}\}$ and $f_i(x^{i,1}) = p(x^{i,1})/\alpha_i$ and $f_i(x^{i,2}) = p(x^{i,2})/\alpha_i$. This completes the proof. □

The proof of Theorem 9 also yields the following:

Corollary 10. *Any distribution p on $\{0,1\}^N$ can be written as a mixture of K elements of the independence model, where K is the minimal number of pairs differing in exactly one entry, which suffices to cover the support of p.*

We now show the optimality of Theorem 9.

Theorem 11. *There exist distributions on $\{0,1\}^N$ which can not be written as a mixture of less than 2^{N-1} elements from the independence model $\overline{\mathcal{E}^1}$. For instance the distributions with support $Z := \{x \in \mathcal{X}\colon \prod_{i\in[N]}(-1)^{x_i} = 1\}$.*

This also means that there exist distributions which cannot be well approximated as a mixture of less than 2^{N-1} elements from $\mathcal{E}^1$.

Proof of Theorem 11.

1. By Proposition 3, the marginal polytope of the independence model on binary vectors of length N is (essentially) the N-cube.
2. The graph of the N-cube is bipartite, or equivalently, 2-colourable.
3. A set is support set of a distribution in the closure of the independence model only if its convex hull is a face of the unit N-cube, Proposition 5.
4. Define Z as the set of vertices of the N-cube which are assigned the same color in a 2-coloring of the graph of the N-cube. Then, $|Z| = 2^N/2$. Furthermore, no subset of Z of cardinality larger than one is facial in the independence model, since otherwise some pair in Z would be an edge of the N-cube, in contradiction to its definition.
5. Consider any probability distribution p with $\mathrm{supp}(p) = Z$.
6. If p is written as a mixture of elements in the closed independence model, $p = \sum_i \alpha_i f_i$, then every f_i (for which $\alpha_i > 0$) must have support contained in Z and this support set must be facial in the independence model, Proposition 2. Hence, $|\mathrm{supp} f_i| = 1, \forall f_i$ for which $\alpha_i > 0$. Therefore, the mixture must have at least $|Z| = \frac{2^N}{2}$ components.
7. $Z := \{x \in \mathcal{X}\colon \prod_{i\in[N]}(-1)^{x_i} = 1\}$ defines a 2-coloring of the N-cube, since for any edge of the N-cube with vertices $\{x^1, x^2\}$ we have that x^1 and x^2 differ in exactly one entry, and thus $\prod_i(-1)^{x_i^1} = -\prod_i(-1)^{x_i^2}$. Clearly, $|Z| = \frac{2^N}{2}$.

□

Now we turn our attention to the decomposability of correlated distributions (say with correlations up to order k) as mixtures of independent distributions.

For the dimension of a mixture of m independent distributions we have from a simple counting argument that $\dim(\mathcal{M}_m^1) \leq mN + m - 1$. We know that the dimension of $\mathcal{E}^k$ is $\dim \mathcal{E}^k = \sum_{j=1}^k \binom{N}{j}$. This gives an easy lower bound for m in order that $\mathcal{M}_m^1 \supseteq \mathcal{E}^k$. Actually, due to the so called dimension defect, the dimension of the mixture model is smaller, but this is a hard problem. However, as can be seen from Theorem 11, a number of parameters $2^{N-1}N + 2^{N-1} - 1 \gg 2^N - 1 = \dim \mathcal{E}^N$ is necessary to cover the simplex, which is $\overline{\mathcal{E}^N}$, (althoug this does not mean that the dimension of a smaller mixture is smaller than $2^N - 1$). The proof of Theorem 11 reveals a way of providing lower bounds for the necessary number of mixture components from the independence model:

Proposition 12. *A necessary condition for the mixture model* $M^1_m := \{p = \sum_{i=1}^m \alpha f_i : f_i \in \overline{\mathcal{E}^1}, \alpha_i \geq 0, \sum \alpha_i = 1\}$ *to contain all distributions from* $\mathcal{E}^k$ *and from* $\overline{\mathcal{E}^k}$ *is that*

$$m \geq \max_{\mathcal{Y},Z}\{|\mathcal{Y}| : \mathcal{Y} \subseteq Z\},$$

where $\mathcal{Y}$ *is* Q_k*-facial, and* Z *is the set of vertices of the* N*-cube with the same color in some* 2*-coloring.*

From Theorem 6, we know that all sets of cardinality $2^k - 1$ are S-sets in $\overline{\mathcal{E}^k}$. If we take a subset of Z, (Z defined e.g. as in Theorem 11), of cardinality $2^k - 1$, then we have from Proposition 12 that in order to have the mixture model containing distributions with support given by that set, we need at least $2^k - 1$ mixture components, i.e. $m \geq 2^k - 1$. This yields:

Corollary 13. *A necessary condition for the mixture model* $M^1_m := \{p = \sum_{i=1}^m \alpha f_i : f_i \in \overline{\mathcal{E}^1}, \alpha_i \geq 0, \sum \alpha_i = 1\}$ *to contain all distributions from* $\mathcal{E}^k$ *and from* $\overline{\mathcal{E}^k}$ *is that*

$$m \geq \max\left\{\frac{1}{N+1}\sum_{j=0}^{k}\binom{N}{j}, 2^k - 1\right\}.$$

It is possible that there are yet other larger simplices which are faces of the marginal polytope, (this is allways the case, see Theorem 16 below), and whose intersection with Z is larger.

For $N = 4$ Q_2 is a polytope of dimension 10. Numerical computations show that many facets (proper faces of maximal dimension, in this case 9) are simplices (which is not surprising regarding that here k is fairly large in relation to N). We found for example that $(0000), (0001), (1100), (1010), (0110), (1001), (0011)$, $(1101), (1011), (1111)$ is a face of dimension 9, (which is a simplex), and its intersection with Z has cardinality 7. For $N = 3$ Q_2 has dimension 6 and we found that $(000), (100), (010)$, $(110), (101), (011)$ is a face which is a simplex with intersection of cardinality 4 with Z. Summarizing:

Corollary 14.

- *A necessary condition for* $\mathcal{M}^1_m$ *to contain* $\overline{\mathcal{E}^2}$ *for* $\mathcal{X} = \{0,1\}^4$ *is that* $m \geq 7$.
- *A necessary condition for* $\mathcal{M}^1_m$ *to contain* $\overline{\mathcal{E}^2}$ *for* $\mathcal{X} = \{0,1\}^3$ *is that* $m \geq 4$.

Mixtures from more general models

Now we turn our attention to mixture decompositions using elements from larger exponential families.

In view of Lemma 7 it is necessary to determine when a face of the marginal polytope is a simplex, and how many such faces suffice to cover all vertices. This problem is related to the problem of optimal covering codes, which is very hard. For example finding a minimum clique cover (partition into cliques) is a graph-theoretical NP-complete problem, or finding perfect covering codes on $\{0,1\}^N$ of general radius is still, in general, an open problem. However, from Kahle's theorem we have that all sets of vertices of a certain cardinality describe faces of the marginal polytope and furthermore that these faces are simplices. An inmediate consequence of Lemma 8 and Theorem 6 is the following:

Theorem 15. (Mixtures of distributions in an exponential family) *Given any $p \in \overline{\mathcal{P}(\mathcal{X})}$, there exist $f_i(\cdot) \in \overline{\mathcal{E}^k(\mathcal{X})}$, and $\alpha_i \geq 0$, $\sum_i \alpha_i = 1$, $i = 1, \dots, \left\lceil \frac{2^N}{2^k-1} \right\rceil$ such that*

$$p(x) = \sum_{i=1}^{\left\lceil \frac{2^N}{2^k-1} \right\rceil} \alpha_i f_i(x).$$

It is possible to derive sharper results using sets larger than $2^k - 1$ (which have the S-property), which in our construction is equivalent to the problem of deriving lower bounds on the number of partition elements beeing S-sets required to cover the sample space. This problem is more difficult because in this case additional structural constraints on the sets arise (in addition to the cardinality constraint). We know that Kahle's result is optimal for $\overline{\mathcal{E}^k}$ in the sense that there exist sets of cardinality 2^k which are not S-sets.

We use the following:

Theorem 16. (Th. 14.4 in Bronsted, Arne (1983)) *Let P be a K-neighborly d-polytope. Then every face F of P with $0 \leq \dim F \leq 2K - 1$ is a simplex, (i.e., P is $2K - 1$-simplicial).*

We know that for $\overline{\mathcal{E}^k}$, the marginal polytope $\mathrm{conv}\{A_y\}_{y\in\mathcal{X}}$ is $(2^k - 1)$-neighborly. This polytope has dimension $|\Delta_k| - 1 = \sum_{i=1}^{k} \binom{N}{i}$. We have therefore by Theorem 16 that all $K := 2(2^k - 1) - 1$-dimensional faces of Q_k are simplices. This yields that all support sets of distributions in $\overline{\mathcal{E}^k}$ of cardinality $2K$, $K = 2^k - 1$ are S-sets. If $2^k - 1 > \left\lfloor \frac{1}{2}(|\Delta_k| - 1) \right\rfloor$, then Q_k is a simplex. It is easy to see that the later only happens when $k = N$.

Since the set of faces $L(P)$ of any convex d-dimensional polytope P is a graded poset (with rank function $r(F) = \dim F + 1$, $F \in L(P)$), (see Theorem 15.1.2 in M. Henk et al. (1997)), we have for any $g \leq d$ that $\cup_{F\in L(P):\dim F = g} F$ contains all vertices of P. This means simply that the union of the support sets of distributions in $\overline{\mathcal{E}^k}$ of cardinality $2K$ is all $\mathcal{X}$.

Now, for the construction of small mixture representations we ask: Given that a d-dimensional polytope P is $2^k - 1$-neighborly, (and K-simplicial, $K = 2(2^k - 1) - 1$), what is the minimal number of K-faces needed to cover all vertices of P?

We will use the following result:

Theorem 17. (M. Develin, 2004) *Suppose P is a d-dimensional polytope which is not the simplex. Then for all $0 < k < d$, we can find a k-face of P and a $(d-k)$-face of P which are disjoint. Equivalently, if P is a d-polytope for which all k-faces of P intersect all $(d-k)$-faces of P, then P must be the simplex.*

We can improve the bound of Theorem 15:

Theorem 18. *Set $\eta = \dim Q_k = \dim \mathcal{E}^k = \sum_{j=1}^{k} \binom{N}{j}$, and $K = 2^k - 1$.*

- *The minimal number of faces of Q_k which are simplices and suffice to cover all vertices is upper bounded by $\frac{2^N}{K} - \left\lfloor \frac{\eta}{2K-1} \right\rfloor + 2$.*
- *For $\mathcal{M}_m^k$ to contain every distribution on $\mathcal{X} = \{0,1\}^N$ it suffices that*

$$m \geq \frac{2^N}{K} - \left\lfloor \frac{\eta}{2K-1} \right\rfloor + 2.$$

This result improves Theorem 15 by a factor between 1 and 2. Especially when N is large, since then η is large compared to K for fixed k. Probably a more detailed analysis allows further improvements.

Proof of Theorem 18. We show that it is possible to decompose the sampling space $\mathcal{X}$ in the specified number of S-sets for the model $\overline{\mathcal{E}^k}$. Then Lemma 8 yields the second statement.

If Q_k is the η-simplex we are done. If not, by Theorem 17 Q_k has a face of dimension $2K-1$ and one face of dimension $\eta - (2K-1)$ which are disjoint. By Theorem 6 and Theorem 16 we have that the $2K-1$-face is a simplex and contains $2K$ vertices. If the disjoint face of complementary dimension is a simplex we are done. If not, it contains at least $\eta - (2K-1) + 2$ vertices and is itself a polytope. Any faces of this polytope are also faces of Q_k. Therefore any of its $2K-1$-faces is a simplex. We use Theorem 17 again on this polytope, and repeat this procedure until the dimension is exhausted. In the worst case no face of dimension larger than $2K-1$ is a simplex and we will get $\left\lfloor \frac{\eta}{2K-1} \right\rfloor$ disjoint faces of dimension $2K-1$ which are simplices and possibly one more disjoint face of smaller dimension which also is a simplex. They cover at least $\left\lfloor \frac{\eta}{2K-1} \right\rfloor 2K$ vertices. All other vertices can be covered by at most $\left\lceil \frac{2^N - \left\lfloor \frac{\eta}{2K-1} \right\rfloor 2K}{K} \right\rceil$ disjoint faces of dimension at most $K-1$ which are simplices, since by Theorem 6 any set of K points or less is a face of the marginal polytope and is a simplex. This completes the proof. □

4 Concluding Remarks

In this paper we did:

- We introduced the concept of S-sets of a model as the regions of the sample space such that all distributions with support therein are contained in that model, (this concept perhaps exists in the literature with another name). We provided a characterization of S-sets of exponential families, Lemma 7, which allows to formulate the mixture decomposition problem as a covering problem for vertices of convex polytopes, Lemma 8. We showed that this formulation provides a meaningful way to decompose distributions as mixtures of elements in the boundary of exponential families and allows to derive relations between the number of mixture components and the representational power of a mixture model.

- For the important class of mixture models where the mixture components belong to the set of independent distributions, we provided a necessary and sufficient relation between the number of binary random variables N and the number of mixture components m to have that the model contains every probability distribution, namely $m = 2^{N-1}$, Theorem 9 and Theorem 11.

- We derived new lower bounds for the number of mixture components from the independence model necessary to represent correlated distributions, Proposition 12 and Corollaries.

- We derived new upper bounds for the minimal number of mixture components from general exponential families necessary to represent any distribution, Theorem 15 and Theorem 18.

Some issues directly related to the work presented here are the following:

- We think the result Theorem 18 can be further improved within our framework. The question is: Let $K = 2^k - 1$. How many $2K - 1$-faces of the marginal polytope Q_k are necessary to cover all its vertices, given that all of them are simplices and that Q_k is K-neighborly? The following observation may be helpful (we do not go into details here): Whenever an S-set of cardinality L exists, then a family of S-sets of cardinality L exists which covers $\mathcal{X}$.

- Theorem 11 indicates that in order to cover a set of distributions using mixtures from the independence model it is required that the mixture elements have S-support. Can we use similar arguments to derive necessary conditions on the number of mixture components from $\overline{\mathcal{E}^k}$ to represent any distribution?

Some interesting observations and questions that arose while writing this paper are:

- Mixture models where the mixture weights are constrained, e.g. belong to some model, are of great interest. We assume that the methods presented in this paper can be used to approach those problems also.

- For $N \geq 2$ the graph of the N-cube has more than $2^{2^{N-2}}$ perfect matchings (sets of disjoint edges covering all vertices). The decomposition used in the proof of Theorem 9 is therefore in general highly non unique. In contrast, the decomposition of distributions with support sets contained in Z is unique, Z beeing the set of points which are assigned the same color in a 2-coloring of the N-cube. Distributions supported in this kind of sets seem to be especially complex. This family of distributions can be used to test the representational power of arbitrary models, for example mixture models with restricted mixture weights.

- The mixture model with m independent mixture components has $Nm + m - 1$ parameters, while the description of all distributions on binary N-vectors requires exactly $2^N - 1$ parameters. Since in order to cover all distributions the mixture model needs $m \geq \frac{2^N}{2}$, we have $N\frac{2^N}{2} + \frac{2^N}{2} - 1$ parameters, which is larger than $2^N - 1$ whenever $N \geq 2$. This expresses the lost arising from the constrained way parameters are used in the mixture model. E.g. a mixture representation is in general non unique. However, when the S-sets of a mixture decomposition are fixed, the number of parameters reduces and the decomposition is unique. Obviously when $\mathcal{M}_m^k = \overline{\mathcal{P}}$, then $\dim \mathrm{M}_m^k = \dim \mathcal{P}$. Can we make statements about the dimension of $\mathcal{M}_m^k$ when $\mathcal{M}_m^k \neq \overline{\mathcal{P}}$?

Acknowledgement: The author wishes to express his gratitude to Nihat Ay.

Appendix

Here we provide the proof of Lemma 7.

Here the matrix $A = (A_{\lambda,x})_{\lambda \in 2^{[N]}, x \in \mathcal{X}}$ is defined by $A_{\lambda,x} = (-1)^{|\{i \in \lambda : x_i = 1\}|}$. We write $A(\Lambda, \mathcal{Y})$ for the submatrix of A consisting of the elements $(A_{\lambda,y})_{\lambda \in \Lambda, y \in \mathcal{Y}}$, and $A(:, x)$ for $A(2^{[N]}, x)$. The rows (columns) of A build an orthonormal basis of $\mathbf{R}^{2^N}$ (A is a Hadammard matrix). The model $\mathcal{E}_\Delta$ is described by $A(\Delta, \mathcal{X})$. We write $\ker A$ for the right kernel of the matrix A, and $\operatorname{rk} A$ for its rank. In addition we writte $\operatorname{supp} p$ for the set of x for which $p(x) \neq 0$. Δ^c denotes the complement of Δ in $2^{[N]}$, and $\mathcal{Y}^c$ the complement of $\mathcal{Y}$ in $\mathcal{X} = \{0, 1\}^N$. We write $\langle \cdot, \cdot \rangle$ for the usual inner product of vectors.

We will use the following description of exponential families, a slightly extension of a result of Geiger et al. (2006) presented by J. Rauh et al. (2009).

Theorem 19. (Geiger et al., 2006, Rauh et al., 2009) *A distribution p is an element of $\overline{\mathcal{E}_\Delta}$ iff p fulfills the equations*

$$p^{m^+} - p^{m^-} = 0 \quad \forall m \in \ker A(\Delta, \mathcal{X}),$$

where $m = m^+ - m^-$, $m^\pm(x) := \max\{0, \pm m(x)\}$ and $p^m := \prod_{x \in \mathcal{X}} (p(x))^{m(x)}$.

We also use the following lemma:

Lemma 20. **(Characterization of Facial Sets, Rauh et al. 2009)** *$\mathcal{Y}$ is facial in $\overline{\mathcal{E}_\Delta}$ iff for any $m \in \ker A(\Delta, \mathcal{X})$ the following holds:* $\mathrm{supp} m^+ \subseteq \mathcal{Y} \Leftrightarrow \mathrm{supp} m^- \subseteq \mathcal{Y}$.

And our last ingredient is the following lemma:

Lemma 21. *Consider any $\Delta \subseteq 2^{[N]}$ and any $\mathcal{Y} \subseteq \mathcal{X}$. The matrix $A(\Delta, \mathcal{Y})$ has full rank* $\min\{|\mathcal{Y}|, |\Delta|\}$ *iff $A(\Delta^c, \mathcal{Y}^c)$ has full rank* $\min\{|\mathcal{Y}^c|, |\Delta^c|\}$.

In particular, $\mathrm{rk}\, A(\Delta^c, \mathcal{Y}^c) = |\Delta^c| \Leftrightarrow \mathrm{rk}\, A(\Delta, \mathcal{Y}) = |\mathcal{Y}|$.

Proof of Lemma 21. 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 we assume that $A(\Delta, \mathcal{Y})$ has full rank $|\Delta|$, to every $z \in \mathcal{Y}^c$ there exists a vector

$$\tilde{v}_z = A(:, z) + \sum_{x \in \mathcal{Y}} \alpha_x A(:, x) \in \mathrm{span}\left\{\{A(:, x)\}_{x \in \mathcal{Y}}, A(:, z)\right\},$$

for which $\tilde{v}_z(\Delta) = (0, \ldots, 0)$ and $\tilde{v}_z(\Delta^c) = v_z$ with some $v_z \in \mathbf{R}^{|\Delta^c|}$. Notice that $2^N = \langle A(:, z), A(:, z)\rangle = \langle \tilde{v}_z, A(:, z)\rangle = \langle v_z, A(\Delta^c, z)\rangle$, since $A(:, z) \perp A(:, x)\ \forall z \neq x$.

For all $y \in \mathcal{Y}^c \setminus \{z\}$ we have that $A(:, y) \perp \mathrm{span}\{\{A(:, x)\}_{x \in \mathcal{Y}}, A(:, z)\}$, and therefore

$$A(\Delta^c, y) \perp v_z \quad \forall z \neq y, \;\; z, y \in \mathcal{Y}^c.$$

Summarizing, there exists a set of vectors $\{v_z\}_{z \in \mathcal{Y}^c}$ s.t.

$$\langle A(\Delta^c, y), v_z\rangle = 2^N \delta_{y,z} \quad \forall y, z \in \mathcal{Y}^c.$$

This can be written as a matrix multiplication:

$$\left[v_{z_1}, \ldots, v_{z_{|\mathcal{Y}^c|}}\right]^\top \cdot A(\Delta^c, \mathcal{Y}^c) = 2^N \,\mathrm{diag}(\mathbb{1}).$$

We have $|\mathcal{Y}^c| = |\Delta^c|$, so that $A(\Delta^c, \mathcal{Y}^c)$ is square. From $\det(A \cdot B) = \det(A) \cdot \det(B)$, $\det A(\Delta^c, \mathcal{Y}^c) \neq 0$ so that it has full rank.

Now consider a $\mathcal{Y}$ for which $|\mathcal{Y}| \neq |\Delta|$. W.l.o.g. $|\mathcal{Y}| \leq |\Delta|$, otherwise use $\mathcal{Y}' = \mathcal{Y}^c$ and $\Delta' = \Delta^c$.

The starting point is $\mathrm{rk}\, A(\Delta, \mathcal{Y}) = |\mathcal{Y}|$. Note that $A(\Delta, \mathcal{X})$ has full rank $|\Delta|$. Therefore, a set $\tilde{\mathcal{Y}}$ exists s.t. $\mathcal{X} \supseteq \tilde{\mathcal{Y}} \supseteq \mathcal{Y}$, $|\tilde{\mathcal{Y}}| = |\Delta|$ and $\mathrm{rk}\, A(\Delta, \tilde{\mathcal{Y}}) = |\Delta|$. From the first part of the proof we have that this is equivalent to $\mathrm{rk}\, A(\Delta^c, \tilde{\mathcal{Y}}^c) = |\Delta^c|$. But this implies $\mathrm{rk}\, A(\Delta^c, \mathcal{Y}^c) = |\Delta^c|$, since $\mathcal{Y}^c \supseteq \tilde{\mathcal{Y}}^c$. For the other direction: $\mathrm{rk}\, A(\Delta^c, \mathcal{Y}^c) = |\Delta^c|$ implies the existence of some $\tilde{\mathcal{Y}}^c \subseteq \mathcal{Y}^c$, $|\tilde{\mathcal{Y}}^c| = |\Delta^c|$ and $\mathrm{rk}\, A(\Delta^c, \tilde{\mathcal{Y}}^c) = |\Delta^c|$. From the first part again, this is $\mathrm{rk}(\Delta, \tilde{\mathcal{Y}}) = |\Delta| = |\tilde{\mathcal{Y}}|$. This again implies $\mathrm{rk}\, A(\Delta, \mathcal{Y}) = |\mathcal{Y}|$, since $\mathcal{Y} \subseteq \tilde{\mathcal{Y}}$. □

Now we come to the proof of Lemma 7:

Proof of Lemma 7. Let $\Delta \subseteq 2^{[N]}$ and $\mathcal{Y} \subseteq \mathcal{X}$. We show that the following statements are equivalent:

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

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

(iii) $\operatorname{supp}(m^+) \cap \mathcal{Y}^c \neq \emptyset$ and $\operatorname{supp}(m^-) \cap \mathcal{Y}^c \neq \emptyset \quad \forall m \in \ker A(\Delta, \mathcal{X})$.

(iv) $\mathcal{Y}$ is facial and $\operatorname{supp}(m^+) \cap \mathcal{Y}^c \neq \emptyset \quad \forall m \in \ker A(\Delta, \mathcal{X})$.

(v) $\mathcal{Y}$ is facial and $\operatorname{supp}(m^-) \cap \mathcal{Y}^c \neq \emptyset \quad \forall m \in \ker A(\Delta, \mathcal{X})$.

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

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

(viii) Every $\mathcal{Y}' \subseteq \mathcal{Y}$ is facial. I.e., $\mathcal{Y}$ corresponds to a face of the marginal polytope which is a simplex.

Item (i) resembles the definition of the S-property for the set $\mathcal{Y}$.

The equivalence of items (iv) and (v) follows inmediatly from Lemma 20. Items (iv) and (v) are both equivalent to item (iii), since by Lemma 20 $\mathcal{Y}^c \cap \operatorname{supp}(m^+) \neq \emptyset \Leftrightarrow \mathcal{Y}^c \cap \operatorname{supp}(m^-) \neq \emptyset$.

The equivalence of items (i) and (ii) reveals that if $\mathcal{Y} \subseteq \mathcal{X}$ has the S property, then every $\mathcal{Y}' \subseteq \mathcal{Y}$ does also. This follows trivially from item (iii), since whenever $\mathcal{Y}^c$ is enlarged, the properties $\operatorname{supp}(m^+) \cap \mathcal{Y}^c \neq \emptyset$ and $\operatorname{supp}(m^-) \cap \mathcal{Y}^c \neq \emptyset$ required there are preserved.

For (ii) *if and only if* (iii): The claim *if* follows directly from Th.19. For the implication *only if* we have to show that if $\operatorname{supp} m^+ \cap \mathcal{Y}^c = \emptyset$ for some $m \in \ker A(\Delta, \mathcal{X})$, then a vector p with support $\mathcal{Y}$ exists which does not satisfy $p^{n^+} - p^{n^-} = 0$ for some $n \in \ker A(\Delta, \mathcal{X})$.

Assume $\operatorname{supp}(m^+) \subseteq \mathcal{Y}$. If there exists any distribution with support $\mathcal{Y}$, then $\mathcal{Y}$ is facial, and by Lemma 20, from $\operatorname{supp} m^+ \subseteq \mathcal{Y}$ we also have that $\operatorname{supp} m^- \subseteq \mathcal{Y}$. Obviously $\operatorname{supp}(m^+) \cap \operatorname{supp}(m^-) = \emptyset$. Assume that some $\tilde{p}$ is contained in $\overline{\mathcal{E}_\Delta}$ and has support $\mathcal{Y}$ (if none exists we are done). For the entries where $m \neq 0$ we write $(\tilde{p}_i)_{\{i:m_i \neq 0\}} = (\tilde{\xi}, \tilde{\eta})$ for some $(\tilde{\xi}, \tilde{\eta}) \in \mathbf{R}_+^{|\operatorname{supp}(m)|}$. If $m \neq \vec{0}$, then $|\operatorname{supp}(m^+)| > 0$ and $|\operatorname{supp}(m^-)| > 0$, since $0 = \langle A(\emptyset, \mathcal{X}), m \rangle = \sum_i m_i$. We can assume that $\|\tilde{\xi}\| < \|\tilde{\eta}\|$, where $\| \cdot \|$ is the sum of the entries of a positive vector, (if not, again we are done). Necessarily we have $\tilde{\xi}^{m^+} - \tilde{\eta}^{m^-} = 0$.

Now consider a vector p which is equal to $\tilde{p}$ in the entries where $m = 0$, and for which $\xi = 2\tilde{\xi}$, and $\eta = (1 - \|\tilde{\xi}\|/\|\tilde{\eta}\|)\tilde{\eta}$. We have $\|\xi\| + \|\eta\| = \|\tilde{\xi}\| + \|\tilde{\eta}\|$, such that p also describes a distribution. For this we have

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

Observe that $\langle \mathbb{1}, m^+\rangle > 0$, and $\langle \mathbb{1}, m^-\rangle > 0$, w.l.o.g. they are both larger than one, since for any $c \in \mathbf{R}$ and $m \in \ker A(\Delta, \mathcal{X})$ we have $c \cdot m \in \ker A(\Delta, \mathcal{X})$. We have also that $0 < \|\tilde{\xi}\|/\|\tilde{\eta}\| < 1$, and ξ_0 is greater than 0 in every entry. Hence, we have the claim.

For (vi) *if and only if* (iv): The statement $\mathcal{Y}^c \cap \operatorname{supp}(m) \neq \emptyset$ coincides with $\mathcal{Y}^c \cap \operatorname{supp}(m^+) \neq \emptyset$ or $\mathcal{Y}^c \cap \operatorname{supp}(m^-) \neq \emptyset$. From Lemma 20 we have that 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 $\mathcal{Y}^c \cap \operatorname{supp}(m) \neq \emptyset \quad \forall m \in \ker A(\Delta, \mathcal{X})$.

Observe that any $m \in \ker A(\Delta, \mathcal{X})$ can be written as

$$m = \sum_{\lambda \in \Delta^c} \alpha_\lambda A(\lambda, \mathcal{X}),$$

which can also be written as $m(x) = \langle \alpha, A(2^{[N]}, x)\rangle$, where $\alpha \in \mathbf{R}^{2^N}$ satisfies $\operatorname{supp}(\alpha) \subseteq \Delta^c$. We have for any $x \in \mathcal{X}$ that

$$m(x) = \langle \alpha, A(2^{[N]}, x)\rangle = 0 \quad \Leftrightarrow \quad \alpha \perp A(2^{[N]}, x).$$

Hence, $\mathcal{Y}^c \cap \operatorname{supp}(m) = \emptyset$, (which is equivalent to $m(x) = 0 \,\forall x \in \mathcal{Y}^c$), is equivalent to the existence of some $\alpha \in \mathbf{R}^{2^N}$ such that

$$\alpha \perp A(2^{[N]}, x) \quad \forall x \in \mathcal{Y}^c. \tag{1}$$

That no $\alpha \neq (0, \ldots, 0)$ with $\operatorname{supp}(\alpha) \subseteq \Delta^c$ can fulfill eqs. 1 is exactly the case when $\operatorname{span}\{A(\Delta^c, x)\}_{x \in \mathcal{Y}^c} = \mathbf{R}^{|\Delta^c|}$, which is equivalent to $\operatorname{rk} A(\Delta^c, \mathcal{Y}^c) = |\Delta^c|$.

Item (vii) is by Lemma 21 equivalent to item (vi).

For (viii) *if and only if* (vii): $\operatorname{rk} A(\Delta, \mathcal{Y}) = |\mathcal{Y}|$ is equivalent to $\{A(\Delta, y)\}_{y \in \mathcal{Y}}$ being linearly independent, from which follows that $\operatorname{conv}\{A(\Delta, y)\}_{y \in \mathcal{Y}}$ is a simplex. Therefore, if $\mathcal{Y}$ is assumed to be facial, *all* sets $\mathcal{Y}' \subseteq \mathcal{Y}$ are facial.

Now assume that $\operatorname{conv}\{A(\Delta, y)\}_{y \in \mathcal{Y}}$ is a face of Q_Δ which is a simplex. Then, $\{A(\Delta, y)\}_{y \in \mathcal{Y}}$ are affine independent. This means that for any $y_0 \in \mathcal{Y}$ the vectors $\{A(\Delta, y) - A(\Delta, y_0)\}_{y \in \mathcal{Y} \setminus \{y_0\}}$ are linearly independent, i.e. the equation

$$\sum_{y \in \mathcal{Y} \setminus \{y_0\}} \beta_y (A(\Delta, y) - A(\Delta, y_0)) = 0 \tag{2}$$

has the unique solution $\beta_y = 0 \,\forall y \in \mathcal{Y} \setminus \{y_0\}$. The equation $\sum_{y \in \mathcal{Y}} \beta_y A(\emptyset, y) = 0$ has only solutions of the form $\beta_{y_0} = \sum_{y \in \mathcal{Y} \setminus \{y_0\}} \beta_y$, since $A(\emptyset, x) = 1 \,\forall x$. Hence, the solution set of $\sum_{y \in \mathcal{Y}} \beta_y A(\Delta, y) = 0$ can be obtained from the solution set of eq. 2. From this the linear independence follows. □

References

Amari, Shun-ichi (1999). Information Geometry on Hierarchical Decomposition of Stochastic Interactions. *IEEE Transaction on Information Theory*, *47*, 1701–1711.

Amari, Shun-ichi (2010). Conditional Mixture Model for Correlated Neural Spikes. *Neural Computation*, *22*, 1718–1736.

Bronsted, Arne (1983). An introduction to convex polytopes. *Springer Verlag*.

Develin, Mike (2004). Disjoint Faces of Complementary Dimension *Contributions to Algebra and Geometrie*, *45* 2, 463–464.

Diaconis, Persi (1977). Finite forms of de Finetti's theorem on exchangeability. *Synthese* vol. 36 2, 271–281.

Drton, Mathias & Sturmfels, Bernd & Sullivant, Seth. Lectures on Algebraic Statistics. *Oberwolfach Seminars vol. 39, Birkhuser*.

Geiger, Dan & Meek, Christopher & Sturmfels, Bernd (2006). On the toric algebra of graphical models. *Ann. Statist.*, *34*, 1463–1492.

Henk, Martin & Richter-Gebert, Jürgen & Ziegler, Günter M. Basic properties of convex polytopes. *Handbook of discrete and computational geometry* pages 243–270.

Kahle, Thomas & Ay, Nihat (2006) Support sets of distributions with given interaction structure. *Proceedings of WUPES'06. Santa Fe Institute Working Paper 06-08-027.*

Kahle, Thomas (2010) Neighborliness of Marginal Polytopes. *Contributions to Algebra and Geometry*, *51* 1, 45–56.

Lindsay, B. G. (1995). Mixture Models: theory, geometry, and applications. *NSF-CBMS Regional Conference Series in Probability and Statistics* Vol 5.

Montufar, Guido & Ay, Nihat (2010). Refinements of Universal Approximation Results for Restricted Boltzmann Machines and Deep Belief Networks. *Submitted to Neural Computation*.

Rauh, Johannes & Kahle, Thomas & Ay, Nihat (2009). Support Sets in Exponential Families and Oriented Matroid Theory. *Proc. WUPES'09, invited for special issue of IJAR*.

Titterington, D.M. & Smith, A. F. M. & Makov, U. E. (1985). Statistical analysis of finite mixture distributions. *John Wiley and Sons*.