Kybernetika

Johannes Rauh

Optimally approximating exponential families

Kybernetika, Vol. 49 (2013), No. 2, 199–215

Persistent URL: http://dml.cz/dmlcz/143362

KYBERNETIKA — VOLUME *49* (2013), NUMBER 2, PAGES 199–215

OPTIMALLY APPROXIMATING EXPONENTIAL FAMILIES

Johannes Rauh

This article studies exponential families $\mathcal{E}$ on finite sets such that the information divergence $D(P\|\mathcal{E})$ of an arbitrary probability distribution from $\mathcal{E}$ is bounded by some constant $D > 0$. A particular class of low-dimensional exponential families that have low values of D can be obtained from partitions of the state space. The main results concern optimality properties of these partition exponential families. The case where $D = \log(2)$ is studied in detail. This case is special, because if $D < \log(2)$, then $\mathcal{E}$ contains all probability measures with full support.

Keywords: exponential family, information divergence

Classification: 94A15, 62B10, 94A17

1. INTRODUCTION

Let $\mathcal{X}$ be a finite set of cardinality N, and denote by $\mathbf{P}(\mathcal{X})$ the set of probability distributions on $\mathcal{X}$. The information divergence $D(P\|Q)$ is a natural distance measure on $\mathbf{P}(\mathcal{X})$. Although $D(P\|Q)$ is not symmetric and does not satisfy the triangle inequality, it is often used by statisticians and information theorists since it appears naturally in many contexts [4, 5]. For any exponential family $\mathcal{E}$ on $\mathcal{X}$ (as defined in Section 2) and any $P \in \mathbf{P}(\mathcal{X})$ write $D_{\mathcal{E}}(P) = \inf_{Q\in\mathcal{E}} D(P\|Q)$.

In 2002 Nihat Ay formulated the problem of finding the maximizers of $D_{\mathcal{E}}$ [1]. See [18] for an overview and further references. The present work builds on recent progress in [19] and [15]. The original motivation was the study of information theoretic learning principles, such as the infomax principle [13] and the IMI principle [2]. Another motivation for studying the function $D_{\mathcal{E}}$ comes from machine learning. From a simplified mathematical perspective, the task of machine learning is to approximate a target distribution by a probability distribution from a given statistical model. The maximum of the function $D_{\mathcal{E}}$ gives a theoretical bound on the learning error, when this error is measured in terms of the information divergence [16].

This article discusses the following question:

- Let $D > 0$, and choose a partial order on the exponential families. Which exponential families are minimal among all exponential families $\mathcal{E}$ satisfying $\max D_{\mathcal{E}} \leq D$? What is the answer to this question under further constraints on $\mathcal{E}$?

There are at least two partial orders of interest:

(i) The partial order induced by the dimensions of the exponential families.

(ii) The partial order by inclusion.

The partial order (i) is particularly important for applications, since the dimension of an exponential family is one of the most important invariants that determine the complexity of all computations. The partial order (ii) can be seen as a "local relaxation": A candidate exponential family $\mathcal{E}$ is only compared to "similar" exponential families, contained in $\mathcal{E}$.

Definition 1.1. Let $\mathcal{H}$ be a set of exponential families. An exponential family $\mathcal{E} \in \mathcal{H}$ is called *inclusion D-optimal among $\mathcal{H}$* for some $D \geq \max D_{\mathcal{E}}$ if every $\mathcal{E}' \in \mathcal{H}$ strictly contained in $\mathcal{E}$ satisfies $\max D_{\mathcal{E}} \leq D < \max D_{\mathcal{E}'}$. An exponential family $\mathcal{E} \in \mathcal{H}$ is called *dimension D-optimal among $\mathcal{H}$* if every exponential family $\mathcal{E}' \in \mathcal{H}$ of smaller dimension satisfies $\max D_{\mathcal{E}} \leq D < \max D_{\mathcal{E}'}$. Exponential families that are inclusion or dimension D-optimal among $\mathcal{H}$ for some D are also called *inclusion* or *dimension optimal among $\mathcal{H}$*, without reference to D. If $\mathcal{H}$ equals the set of all exponential families, then the reference to $\mathcal{H}$ may be omitted in all definitions.

The understanding of dimension optimality can be summarized in the constant

$$D_{N,k}(\mathcal{H}) = \min \{\max D_{\mathcal{E}} : \mathcal{E} \in \mathcal{H} \text{ is an exponential family of dimension } k \text{ on } [N]\} .$$

As an example, the set $\mathcal{H}$ may be the set of hierarchical models, the set of graphical models or the set $\mathcal{H}_{\mathbf{1}}$ of exponential families containing the uniform distribution. Obviously, any dimension optimal model is also inclusion optimal. The converse statement does not hold, see Example 5.2 below.

A D-optimal exponential family $\mathcal{E}$ can approximate arbitrary probability measures well, up to a maximal divergence of D. Yaroslav Bulatov proposed to use such exponential families in machine learning (personal communication), for example when using the *minimax algorithm* [21] by Zhu, Wu and Mumford or the *feature induction algorithm* [7] by Della Pietra, Della Pietra and Lafferty. Both algorithms inductively construct an exponential family by adding functions ("features") to the tangent space in order to approximate a given distribution. Applications of the results of the present paper to machine learning will not be discussed here, but in a future work.

One motivation to restrict the class $\mathcal{H}$ of exponential families is that the learning system may not be able to represent arbitrary exponential families. Another motivation is given by Jaynes' principle of maximum entropy [10], which suggests to use the class $\mathcal{H}_{\mathbf{1}}$ of exponential families with uniform reference measure.

This paper also introduces the class of *partition models* (see Section 3): A probability measure P belongs to the partition model associated to a partition $\mathcal{X}' = (\mathcal{X}^1, \dots, \mathcal{X}^{N'})$ if the restriction of P to each block $\mathcal{X}^i$ is uniform. Conjecture 5.4 relates partition models to the above question:

Conjecture 5.4 $D_{N,k} = \log\lceil \frac{N}{k+1} \rceil$, and the dimension $D_{N,k}$-optimal exponential families containing the uniform distribution are partition models.

The results in Section 4 show that the conjecture is true if $\lceil\frac{N}{k+1}\rceil \leq 2$. Theorem 5.3 states that the conjecture is true if $k+1$ divides N, and that, in general, the inequality $D_{N,k} \geq \log\frac{N}{k+1}$ holds.

This paper is organized as follows: Section 2 collects the necessary preliminaries about exponential families and the information divergence. Section 3 introduces partition models and studies their basic properties. log(2)-optimal exponential families $\mathcal{E}$ are studied in Section 4. Section 5 presents results on D-optimal exponential families for arbitrary D.

2. PRELIMINARIES

This section collects known facts that are needed in later sections. It starts with some notions from matroid theory before defining exponential families, the information divergence and hierarchical models. The last part discusses the function $\overline{D}_{\mathcal{E}}$, which arises naturally when studying the maximizers of $D_{\mathcal{E}}$.

2.1. Circuits

In this paper only representable matroids will play a role, but nevertheless the language of abstract matroids is useful. See [17] for an introduction.

Definition 2.1. Let $\mathcal{N}$ be a linear subspace of $\mathbb{R}^{\mathcal{X}}$. The *support* of $u \in \mathcal{N}$ is defined as $\operatorname{supp}(u) := \{x \in \mathcal{X} : u(x) \neq 0\}$. A vector $v \in \mathcal{N} \setminus \{0\}$ is called a *circuit vector* if and only if for any $u \in \mathcal{N}$ satisfying $\operatorname{supp}(u) \subseteq \operatorname{supp}(v)$ there exists $\alpha \in \mathbb{R}$ such that $u = \alpha v$. In other words, circuit vectors are vectors with minimal support. The support $\operatorname{supp}(u)$ of a circuit vector u is called a *circuit*. A finite set $\mathcal{C} \subseteq \mathcal{N}$ is a *circuit basis* if and only if the map $u \in \mathcal{C} \mapsto \operatorname{supp}(u)$ is injective and maps onto the set of circuits.

Lemma 2.2. For every nonzero vector $u \in \mathcal{N}$ and any $x \in \mathcal{X}$ such that $u(x) \neq 0$ there exists a circuit vector $c \in \mathcal{N}$ such that $\operatorname{supp}(c) \subseteq \operatorname{supp}(u)$ and $c(x) \neq 0$.

Proof. Let c be a vector with inclusion-minimal support that satisfies $\operatorname{supp}(c) \subseteq \operatorname{supp}(u)$ and $c(x) \neq 0$. If c is not a circuit vector, then there exists a circuit vector c' with $\operatorname{supp}(c') \subset \operatorname{supp}(c)$. A suitable linear combination $c + \alpha c'$, $\alpha \in \mathbb{R}$ gives a contradiction to the minimality of c. $\square$

It follows that any circuit basis of $\mathcal{N}$ contains a spanning set.

2.2. Exponential families and the information divergence

In this work only exponential families on a finite set $\mathcal{X}$ are studied, for the information divergence from a finite-dimensional exponential family on an infinite set is usually unbounded, cf. Theorem 5.3. See [3] and [5] for an introduction to exponential families and the information divergence.

Let $\tilde{\mathcal{T}}$ be a linear subspace of $\mathbb{R}^{\mathcal{X}}$ containing the constant functions, and let ν be a strictly positive measure on $\mathcal{X}$. The set $\mathcal{E} = \mathcal{E}_{\nu,\tilde{\mathcal{T}}}$ of all probability measures on $\mathcal{X}$ of the form

$$P_\vartheta(x) = \frac{\nu(x)}{Z_\vartheta} e^{\vartheta(x)}$$

is called an *exponential family*. ν is a *reference measure*, and $\tilde{\mathcal{T}}$ will be called the *extended tangent space* of $\mathcal{E}$. The extended tangent space carries its name since its image modulo the constant functions is isomorphic to the tangent space of the manifold $\mathcal{E}$ at any point. The orthogonal complement $\mathcal{N} := \tilde{\mathcal{T}}^\perp$ will be called the *normal space* of $\mathcal{E}$. The normal space is orthogonal to the tangent space of $\mathcal{E}$ at any point $P \in \mathcal{E}$ with respect to the Fisher metric at P.

The exponential family $\mathcal{E}_{\nu,\tilde{\mathcal{T}}}$ can be parametrized as follows: If $a_1, \dots, a_h \in \mathbb{R}^{\mathcal{X}}$ form a spanning set of $\tilde{\mathcal{T}}$, then $\mathcal{E}$ consists of all probability distributions of the form

$$P_\theta(x) = \frac{\nu(x)}{Z_\theta} \exp\left(\sum_{i=1}^{h} \theta_i a_i(x)\right).$$

In this formula $\theta \in \mathbb{R}^h$ is a vector of parameters, and Z_θ ensures normalization. The matrix $A = (a_i(x))_{i,x} \in \mathbb{R}^{h\times\mathcal{X}}$ is called a *sufficient statistics* of $\mathcal{E}$. The linear map corresponding to A is called the *moment map*, denoted by π_A. The columns of A will be denoted by $A_x, x \in \mathcal{X}$. The normal space of $\mathcal{E}$ equals $\mathcal{N} = \{u \in \ker A : \sum_x u(x) = 0\}$. The convex hull of $\{A_x : x \in \mathcal{X}\}$ is a polytope called the *convex support* $\mathbf{M}_A$ of $\mathcal{E}$. This polytope is independent of the choice of $a_1, \dots, a_h$ up to an affine transformation.

Any function $u \in \mathbb{R}^{\mathcal{X}}$ can be decomposed uniquely as a difference $u = u^+ - u^-$ of non-negative functions such that $\operatorname{supp}(u^+) \cap \operatorname{supp}(u^-) = \emptyset$. The following implicit description of an exponential family is useful in many contexts.

Theorem 2.3. Let $\mathcal{E}$ be an exponential family with normal space $\mathcal{N}$ and reference measure ν, and let $\mathcal{C}$ be a circuit basis of $\mathcal{N}$. A probability measure P on $\mathcal{X}$ belongs to $\overline{\mathcal{E}}$ if and only if P satisfies

$$\prod_{x\in\mathcal{X}} \left(\frac{P(x)}{\nu(x)}\right)^{u^+(x)} = \prod_{x\in\mathcal{X}} \left(\frac{P(x)}{\nu(x)}\right)^{u^-(x)}, \qquad \text{for all } u = u^+ - u^- \in \mathcal{C}.$$

Proof. See [20, Theorem 10]. □

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

$$\left\{P = \sum_{i=1}^{c} \lambda_i P_i : P_1 \in \mathcal{E}_1, \dots, P_c \in \mathcal{E}_c \text{ and } \lambda_1 \geq 0, \dots, \lambda_c \geq 0, \sum_{i=1}^{c} \lambda_i = 1\right\}.$$

The topological closure of $\mathcal{E}$ will be denoted by $\overline{\mathcal{E}}$, and $\mathbf{P}(\mathcal{X})^\circ$ denotes the interior of $\mathbf{P}(\mathcal{X})$, which consists of all probability measures with full support.

Corollary 2.4. Let $\mathcal{E}$ be an exponential family with normal space $\mathcal{N}$. Let $\mathcal{Y} \subset \mathcal{X}$. If every circuit vector $c \in \mathcal{N}$ satisfies $\mathrm{supp}(c) \subseteq \mathcal{Y}$ or $\mathrm{supp}(c) \subseteq \mathcal{X} \setminus \mathcal{Y}$, then $\overline{\mathcal{E}}$ equals the mixture of $\overline{\mathcal{E}} \cap \mathbf{P}(\mathcal{Y})$ and $\overline{\mathcal{E}} \cap \mathbf{P}(\mathcal{X} \setminus \mathcal{Y})$.

Proof. For any $P \in \mathbf{P}(\mathcal{X})$ and $\mathcal{Y} \subseteq \mathcal{X}$ with $P(\mathcal{Y}) > 0$ define the *truncation*

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

By Theorem 2.3, a probability measure $P \in \mathbf{P}(\mathcal{X})$ with full support lies in $\mathcal{E}$ if and only if its truncations $P^{\mathcal{Y}}$ and $P^{\mathcal{X} \setminus \mathcal{Y}}$ lie in $\mathcal{E} \cap \mathbf{P}(\mathcal{Y})$ and $\mathcal{E} \cap \mathbf{P}(\mathcal{X} \setminus \mathcal{Y})$, respectively. □

In the language of matroid theory, the corollary says the following: If $\mathcal{X}_1, \dots, \mathcal{X}_c$ are the connected components of the matroid of $\mathcal{N}$, then $\overline{\mathcal{E}}$ equals the mixture of $\overline{\mathcal{E}_1}, \dots, \overline{\mathcal{E}_c}$, where $\mathcal{E}_i = \overline{\mathcal{E}} \cap \mathbf{P}(\mathcal{X}_i)^\circ$ is an exponential family on $\mathcal{X}_i$ for $i = 1, \dots, c$.

The *information divergence* (also known as the *Kullback-Leibler divergence* or *relative entropy*) of positive measures P, Q is defined as

$$D(P\|Q) = \sum_{x \in \mathcal{X}} P(x) \log \left(\frac{P(x)}{Q(x)} \right)$$

with the convention that $0 \log 0 = 0 \log(0/0) = 0$. It is finite unless $\mathrm{supp}(P)$ is not contained in $\mathrm{supp}(Q)$. If ν equals the counting measure on $\mathcal{X}$ (i. e. $\nu_x = 1$ for all x), then $D(P\|\nu)$ equals minus the Shannon entropy $H(P)$. If P and Q are probability measures, then $D(P\|Q)$ is strictly positive unless $P = Q$.

Let $\mathcal{E}$ be an exponential family. For any probability measure P on $\mathcal{X}$ there is a unique probability distribution $P_{\mathcal{E}} \in \overline{\mathcal{E}}$ such that $D(P\|P_{\mathcal{E}}) = \inf_{Q \in \mathcal{E}} D(P\|Q)$, see [6]. The measure $P_{\mathcal{E}}$ is called the *(generalized) rI-projection* of P to $\mathcal{E}$ or the *(generalized) MLE.* It can also be characterized as the unique probability measure $P_{\mathcal{E}} \in \overline{\mathcal{E}}$ such that $AP = AP_{\mathcal{E}}$. Alternatively, $P_{\mathcal{E}}$ minimizes the function $D(Q\|\nu)$ on $\{Q \in \mathbf{P}(\mathcal{X}) : P - Q \in \mathcal{N}\}$. In particular, if ν is the counting measure, then $P_{\mathcal{E}}$ maximizes the entropy.

The *Pythagorean theorem* of information theory says that

$$D(P\|\nu) = D(P\|P_{\mathcal{E}}) + D(P_{\mathcal{E}}\|\nu),$$

whenever ν is a reference measure of $\mathcal{E}$. In particular, if $\mathcal{E}$ contains the uniform distribution, then $D_{\mathcal{E}}(P) = H(P_{\mathcal{E}}) - H(P)$.

2.3. Hierarchical loglinear models

Let $\mathcal{X}_1, \dots, \mathcal{X}_n$ be finite sets of cardinality $|\mathcal{X}_i| = N_i$, and let $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$. For any subset $S \subseteq [n]$ let $\mathcal{X}_S = \times_{i \in S} \mathcal{X}_i$. The restrictions $X_i : \mathcal{X} \to \mathcal{X}_i$ to the subsystems can be viewed as random variables, and hierarchical models can be used to study the relationship of these discrete random variables. This section summarizes the main facts which are needed in the following. See [12] and [8] for further information.

Definition 2.5. For any family Δ of subsets of $[n]$ let $\mathcal{E}'_\Delta$ be the set of all probability measures $P \in \mathbf{P}(\mathcal{X})$ that can be written in the form

$$P(x) = \prod_{S \in \Delta} f_S(x), \tag{1}$$

where each f_S is a non-negative function on $\mathcal{X}$ that depends only on those components of x lying in S. In other words, $f_S(x) = f_S(y)$ for all $x = (x_i)_{i=1}^n, y = (y_i)_{i=1}^n \in \mathcal{X}$ satisfying $x_i = y_i$ for all $i \in S$. The *hierarchical exponential family* $\mathcal{E}_\Delta$ of Δ is defined as $\mathcal{E}'_\Delta \cap \mathbf{P}(\mathcal{X})^\circ$. The closure of $\mathcal{E}_\Delta$ (which equals the closure of $\mathcal{E}'_\Delta$) is called the *hierarchical model* of Δ.

In general, $\mathcal{E}'_\Delta \neq \overline{\mathcal{E}_\Delta}$, see [9]. Under some circumstances, when the factorizability probability is important, one might want to call $\mathcal{E}'_\Delta$ a hierarchical model. When studying optimization problems it is more important that the models are closed.

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

It is easy to see that hierarchical exponential families are indeed exponential families: Namely, (1) implies that $\mathcal{E}_\Delta$ consists of all $P \in \mathbf{P}(\mathcal{X})^\circ$ that satisfy

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

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

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

induce the moment map

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

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

Lemma 2.6. Let Δ be a collection of subsets of $[n]$, and let $K = \cup_{J \in \Delta} J$. The marginal polytope of Δ is (affinely equivalent to) a 0-1-polytope with $\prod_{i \in K} N_i$ vertices.

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

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

The function $D_{\mathcal{E}}$ is related to the function

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

defined on $\mathcal{N}$. It turns out that the local maximizers of the two functions $D_{\mathcal{E}}$ and $\overline{D}_{\mathcal{E}}$ are in one-to-one correspondence. In fact, this bijection extends to the set of critical points. This correspondence is explained in detail in [15, 18, 19].

The function $\overline{D}_{\mathcal{E}}$ satisfies $\overline{D}_{\mathcal{E}}(\alpha u) = \alpha \overline{D}_{\mathcal{E}}$ for all $\alpha \in \mathbb{R}$ and $u \in \mathcal{N}$. It will mostly be considered on a subset $\partial \mathbf{U}_{\mathcal{N}}$ of $\mathcal{N}$, defined as follows:

Definition 2.7. For any $v \in \mathbb{R}^{\mathcal{X}}$ and $\mathcal{Z} \subseteq \mathcal{X}$ write $v(\mathcal{Z}) := \sum_{x \in \mathcal{Z}} v(x)$. Let

$$\partial \mathbf{U}_{\mathcal{N}} := \left\{ u \in \mathcal{N} : u^+(\mathcal{X}) = u^-(\mathcal{X}) = 1 \right\}.$$

The map $\Psi^+ : u \mapsto u^+$ maps $\partial \mathbf{U}_{\mathcal{N}}$ to a subset of $\mathbf{P}(\mathcal{X})$. A probability distribution in the image of Ψ^+ is called a *kernel distribution*.

In the other direction there is the natural map $\Psi_{\mathcal{E}} : \mathbf{P}(\mathcal{X}) \setminus \overline{\mathcal{E}} \to \mathcal{N}$, defined via

$$\Psi_{\mathcal{E}}(P) = \frac{P - P_{\mathcal{E}}}{(P - P_{\mathcal{E}})^+(\mathcal{X})}.$$

The denominator makes sure that the image of $\Psi_{\mathcal{E}}$ lies in $\partial \mathbf{U}_{\mathcal{N}}$. Since $P = P_{\mathcal{E}}$ if and only if $P \in \overline{\mathcal{E}}$, the map is well-defined on $\mathbf{P}(\mathcal{X}) \setminus \overline{\mathcal{E}}$.

Theorem 2.8. Let $\mathcal{E}$ be an exponential family with normal space $\mathcal{N} \neq 0$. The map $\Psi_{\mathcal{E}}$ restricts to a bijection from the set of local maximizers of $D_{\mathcal{E}}$ to the set of local maximizers of $\overline{D}_{\mathcal{E}}$. An inverse is given by the restriction of the map $\Psi^+ : u \mapsto u^+$. If $P \in \mathbf{P}(\mathcal{X})$ and $u \in \partial \mathbf{U}_{\mathcal{N}}$ are local maximizers of $D_{\mathcal{E}}$ and $\overline{D}_{\mathcal{E}}$, respectively, then

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

Proof. See [15, Theorem 1]. □

Corollary 2.9. Let $\mathcal{E}$ be an exponential family of codimension one. Then $\mathcal{N}$ is spanned by a single vector. This vector can be chosen to be of the form $u = u^+ - u^-$, where u^+ and u^- are two probability distributions of disjoint support. Then u^+ and u^- are the only local maximizers of $D_{\mathcal{E}}$.

Proof. Observe that under the assumptions of the corollary $\partial \mathbf{U}_{\mathcal{N}} = \{+u, -u\}$ is a discrete space. Hence, $\overline{D}_{\mathcal{E}}$ has precisely two local maximizers, corresponding to the two local maximizers u^+ and u^- of $D_{\mathcal{E}}$. □

Corollary 2.10. Let $\mathcal{E}$ be an exponential family. If $\overline{\mathcal{E}} \neq \mathbf{P}(\mathcal{X})$, then $\max D_{\mathcal{E}} \geq \log(2)$.

P r o o f. Let $u \in \partial\mathbf{U}_{\mathcal{N}}$ be a global maximizer of $\overline{D}_{\mathcal{E}}$. Since $\overline{D}_{\mathcal{E}}(-u) = -\overline{D}_{\mathcal{E}}(u)$ the maximal value $\overline{D}_{\mathcal{E}}(u)$ is non-negative. Hence $D_{\mathcal{E}}(u^+) = \log(1 + \exp(\overline{D}_{\mathcal{E}}(u))) \geq \log(2)$. □

It is straightforward to compute the first-order criticality conditions of $\overline{D}_{\mathcal{E}}$:

Proposition 2.11. Let $\mathcal{E}$ be an exponential family with normal space $\mathcal{N}$, let $u \in \partial\mathbf{U}_{\mathcal{N}}$ be a local maximizer of $\overline{D}_{\mathcal{E}}$, and let $\mathcal{Y} = \mathrm{supp}(u)$. The following statements hold:

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

(*ii*) Let $P_{\mathcal{E}}$ be the rI-projection of u^+ and u^-, and let $v \in \mathcal{N}$. Then

$$\sum_{x \in \mathcal{X} \setminus \mathcal{Y}} v(x) \log \frac{|v(x)|}{\nu_x} \leq v^+(\mathcal{Z}')\overline{D}_{\mathcal{E}}(v_0).$$

P r o o f. See [15] or [18, Proposition 3.21]. □

3. PARTITION MODELS

Partition exponential families are convex exponential families. The information divergence from convex exponential families has been studied in [14]. Apart from this, partition exponential families do not seem to have been studied before, despite their peculiar properties. In other contexts the name "partition model" is used for other mathematical objects, but there seems to be little danger of confusion.

Definition 3.1. A *partition* $\mathcal{X}'$ of $\mathcal{X}$ is a family $\mathcal{X}' = \left\{\mathcal{X}^1, \mathcal{X}^2, \dots, \mathcal{X}^{N'}\right\}$ of nonempty subsets $\mathcal{X}^i \subset \mathcal{X}$ such that $\mathcal{X} = \mathcal{X}^1 \cup \mathcal{X}^2 \cup \dots \cup \mathcal{X}^{N'}$ and $\mathcal{X}^i \cap \mathcal{X}^j = \emptyset$ for all $1 \leq i < j \leq N'$. The subsets $\mathcal{X}^i \subseteq \mathcal{X}$ are called the *blocks* of the partition $\mathcal{X}'$. For any $x \in \mathcal{X}$ the block $\mathcal{X}^i$ containing x is denoted by $\mathcal{X}^x$.

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

Definition 3.2. Let $\mathcal{X}'$ be a partition of $\mathcal{X}$. Denote $\mathbb{R}^{\mathcal{X}'}$ the set of functions $\vartheta : \mathcal{X} \to \mathbb{R}$ such that $x \sim_{\mathcal{X}'} y$ implies $\vartheta(x) = \vartheta(y)$. The exponential family $\mathcal{E}_{\mathcal{X}'}$ with uniform reference measure and extended tangent space $\mathbb{R}^{\mathcal{X}'}$ is called the *partition exponential family* of $\mathcal{X}'$, and $\overline{\mathcal{E}_{\mathcal{X}'}}$ is the *partition model* of $\mathcal{X}'$.

Partition models are, in fact, also linear families: $\overline{\mathcal{E}_{\mathcal{X}'}}$ equals the intersection of $\mathbf{P}(\mathcal{X})$ with the linear space $\mathbb{R}^{\mathcal{X}'}$. In particular, partition exponential families are convex. Convex exponential families have been studied in [14], which contains more detailed arguments for the following calculations. It follows from [14, Proposition 1] that a convex exponential family is a partition exponential family if and only if it contains the uniform distribution.

Remark 3.3. Partition models can be used to model symmetries, as noted in [11]. If a symmetry group G acts on $\mathcal{X}$, then it induces a partition $\mathcal{X}^G$ of $\mathcal{X}$ into orbits $\mathcal{X}^1, \dots, \mathcal{X}^{N'}$. The action of G extends naturally to an action on $\mathbb{R}^{\mathcal{X}}$. Any exponential family that consists of G-invariant probability measures is a subfamily of $\mathcal{E}_{\mathcal{X}^G}$ (such exponential families are called *G-exchangeable* in [11]). Conversely, an arbitrary partition model $\overline{\mathcal{E}_{\mathcal{X}'}}$ arises in this way from the group of all permutations g of $\mathcal{X}$ such that $g(\mathcal{X}^i) = \mathcal{X}^i$ for all $\mathcal{X}^i \in \mathcal{X}'$.

Lemma 3.4. An exponential family with uniform reference measure and sufficient statistics $A \in \mathbb{R}^{h\times\mathcal{X}}$ is a partition exponential family if and only if its convex support is a simplex with vertex set $\{A_x : x \in \mathcal{X}\}$.

Proof. A sufficient statistics of $\mathcal{E}_{\mathcal{X}'}$ is given by the characteristic functions $a_i = \mathbf{1}_{\mathcal{X}^i}$ of the blocks of $\mathcal{X}'$. Any column of $A = (a_i(x))_{i,x}$ is a unit vector, and therefore the convex support is a simplex.

In the other direction define an equivalence relation $\sim$ on $\mathcal{X}$ via $x \sim y$ if and only if $A_x = A_y$. Then $\mathcal{E}$ agrees with the partition exponential family of this equivalence relation. □

For partition models the mapping $P \mapsto P_{\mathcal{E}}$ is easy to compute: The equation $AP = AP_{\mathcal{E}}$ translates into $P(\mathcal{X}^i) = P_{\mathcal{E}}(\mathcal{X}^i)$ for $i = 1, \dots, N'$. Therefore,

$$P_{\mathcal{E}}(x) = P_{\mathcal{E}}^{\mathcal{X}^x}(x)P(\mathcal{X}^x), \qquad \text{for all } x \in \mathcal{X}, \tag{2}$$

where $P_{\mathcal{E}}^{\mathcal{X}^x}$ denotes the truncation of $P_{\mathcal{E}}$ to $\mathcal{X}^x$. Since $P_{\mathcal{E}}$ maximizes the entropy subject to (2), it follows that $P_{\mathcal{E}}^{\mathcal{X}^x} = \frac{1}{|\mathcal{X}^x|}\mathbf{1}_{\mathcal{X}^x}$ is the uniform distribution on $\mathcal{X}^x$. Hence the rI-projection map $P \mapsto P_{\mathcal{E}}$ averages over the blocks of the partition. It follows that

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

As a consequence:

Lemma 3.5. If $\overline{\mathcal{E}}$ is a partition model of a partition $\{\mathcal{X}^1, \dots, \mathcal{X}^{N'}\}$ of coarseness c, then $\max D_{\mathcal{E}} = \log(c)$. A probability measure $P \in \mathbf{P}(\mathcal{X})$ maximizes $D_{\mathcal{E}}$ if and only if the following two conditions are satisfied:

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

(*ii*) $P^{\mathcal{X}^i}$ is a point measure for all i such that $|\mathcal{X}^i| = c$ and $P(\mathcal{X}^i) > 0$.

Corollary 3.6. Let $\overline{\mathcal{E}}$ be the partition model of a partition $\mathcal{X}'$ of coarseness c, and let $\mathcal{Z}$ be the union of the blocks of $\mathcal{X}'$ of cardinality c. Then any $Q \in \overline{\mathcal{E}}$ with support contained in $\mathcal{Z}$ is the rI-projection of some global maximizer of $D_{\mathcal{E}}$. In particular, if $\mathcal{X}'$ is homogeneous, then any $Q \in \overline{\mathcal{E}}$ is the rI-projection of some global maximizer of $D_{\mathcal{E}}$.

Proof. For any $\mathcal{X}^i \in \mathcal{X}'$ of cardinality c choose a representative $x_i \in \mathcal{X}^i$. Define $P \in \mathbf{P}(\mathcal{X})$ by $P(\mathcal{X}^i) = Q(\mathcal{X}^i)$ and $P^{\mathcal{X}^i} = \delta_{x_i}$ for all i such that $|\mathcal{X}^i| = c$. Then $P_{\mathcal{E}} = Q$, so the statement follows from Lemma 3.5. □

Remark 3.7. Composite systems have natural homogeneous partitions, which lead to hierarchical models as defined in Section 2: Suppose that $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$, and let $K \subseteq \{1, \dots, n\}$. Define an equivalence relation $\sim_K$ on $\mathcal{X}$ via $x \sim_K y$ if and only if $x_i = y_i$ for all $i \in K$. The equivalence classes of $\sim_K$ form a homogeneous partition of $\mathcal{X}$ of coarseness $\prod_{i:i\notin K} N_i$. The corresponding partition model $\mathcal{E}_K$ equals the hierarchical exponential family $\mathcal{E}_{\{K\}}$. Conversely, any homogeneous partition $\mathcal{X}'$ can be used to find a bijection of $\mathcal{X}$ with a composite system $\mathcal{X}_1 \times \mathcal{X}_2$, where $\mathcal{X}_1 = \mathcal{X}'$ and $\mathcal{X}_2 \in \mathcal{X}'$. Then the partition $\mathcal{X}'$ arises from $\sim_K$, where $K = \{1\}$.

4. EXPONENTIAL FAMILIES WITH $\max D_{\mathcal{E}} = \log(2)$

By Corollary 2.10 the maximal value of $D_{\mathcal{E}}$ is at least $\log(2)$ unless $\overline{\mathcal{E}} = \mathbf{P}(\mathcal{X})$. This section studies exponential families $\mathcal{E}$ where $\max D_{\mathcal{E}} = \log(2)$. For such an exponential family, any kernel distribution is a local maximizer of $D_{\mathcal{E}}$. Furthermore, $\overline{D}_{\mathcal{E}}(u) = 0$ for all $u \in \mathcal{N}$ (even if $u \notin \partial\mathbf{U}_{\mathcal{N}}$). The main results are:

Theorem 4.1. Let $\mathcal{E}$ be an exponential family on a finite set $\mathcal{X}$ of cardinality N. If $\max D_{\mathcal{E}} = \log(2)$, then the dimension of $\mathcal{E}$ is at least $\lceil \frac{N}{2} \rceil - 1$.

Theorem 4.2. Let $\mathcal{X}$ be a finite set of cardinality N, and let $\mathcal{E}$ be an exponential family on $\mathcal{X}$ of dimension $\lceil \frac{N}{2} \rceil - 1$ satisfying $\max D_{\mathcal{E}} = \log(2)$. If N is even, then $\mathcal{E}$ is a partition model. If N is odd, then there is a set $\mathcal{Z} \subseteq \mathcal{X}$ of cardinality three, a partition model $\overline{\mathcal{E}_{\mathcal{X}\setminus\mathcal{Z}}}$ on $\mathcal{X} \setminus \mathcal{Z}$ and a one-dimensional exponential family $\mathcal{E}_{\mathcal{Z}}$ on $\mathcal{Z}$ such that $\max D(\cdot\|\mathcal{E}_{\mathcal{X}\setminus\mathcal{Z}}) = \log(2) = \max D(\cdot\|\mathcal{E}_{\mathcal{Z}})$, and the closure $\overline{\mathcal{E}}$ equals the mixture of $\overline{\mathcal{E}_{\mathcal{X}\setminus\mathcal{Z}}}$ and $\overline{\mathcal{E}_{\mathcal{Z}}}$. If $\mathcal{E}$ contains the uniform distribution, then $\overline{\mathcal{E}}$ is a partition model.

Proposition 4.3. Let $\mathcal{X} = \{1, 2, 3\}$. For any $u \in \mathbb{R}^{\mathcal{X}}$ such that $u_1 + u_2 + u_3 = 0$ there exists a unique exponential family $\mathcal{E}$ on $\mathcal{X}$ with normal space $\mathcal{N} = \mathbb{R}u$ such that $\max D_{\mathcal{E}} = \log(2)$.

The proofs of the three results will be given below after a series of preliminary lemmas. Under the additional assumptions that N is even Theorem 4.2 has a simpler proof, see Theorem 5.3.

Let $\mathcal{E}$ be an exponential family with sufficient statistics A and normal space $\mathcal{N}$.

Lemma 4.4. For any $v_0, v_1, \dots, v_s \in \mathcal{N}$ let $\mathcal{Z} = \operatorname{supp}(v_0) \setminus \cup_{j=1}^s \operatorname{supp}(v_j)$. Suppose that $\max D_{\mathcal{E}} = \log(2)$. Then

$$\sum_{x\in\mathcal{Z}} v(x) \log \frac{|v(x)|}{\nu_x} = 0 \quad \text{and} \quad \sum_{x\in\mathcal{Z}} v(x) = 0 \quad \text{for all } v \in \mathcal{N}.$$

Proof. The proof is by induction on s. Let $s = 0$. Any $v_0 \in \mathcal{N}$ satisfies $\overline{D}_{\mathcal{E}}(v_0) = 0$ and is a local maximizer of $\overline{D}_{\mathcal{E}}$. The equality $v(\mathcal{Z}) = 0$ for all $v \in \mathcal{N}$ follows from Proposition 2.11 (i). Let $\mathcal{Z}' = \mathcal{X} \setminus \mathcal{Z}$. By Proposition 2.11 ($ii$),

$$\sum_{x\in\mathcal{Z}'} v(x) \log \frac{|v(x)|}{\nu_x} \le v^+(\mathcal{Z}')\overline{D}_{\mathcal{E}}(v_0) = 0$$

for all $v \in \mathcal{N}$. Together with the same inequality with v replaced by $-v$ this implies $\sum_{x\in\mathcal{Z}'} v(x)\log\frac{|v(x)|}{\nu_x} = 0$. Therefore $\sum_{x\in\mathcal{Z}} v(x)\log\frac{|v(x)|}{\nu_x} = \overline{D}_{\mathcal{E}}(v) - \sum_{x\in\mathcal{Z}'} v(x)\log\frac{|v(x)|}{\nu_x}$ vanishes.

If $s \geq 1$, then let $\mathcal{Y} = \mathcal{X} \setminus \operatorname{supp}(v_s)$. Let $\mathcal{E}'$ be the exponential family on $\mathcal{Y}$ with reference measure the restriction $\nu|_{\mathcal{Y}}$ of ν to $\mathcal{Y}$ and normal space $\mathcal{N}' = \{v|_{\mathcal{Y}} : v \in \mathcal{N}\}$. The case $s = 0$ implies $\overline{D}_{\mathcal{E}'}(w) = \overline{D}_{\mathcal{E}}(v) - \sum_{x\in\operatorname{supp}(v_s)} v(x)\log\frac{|v(x)|}{\nu_x} = 0$ for all $w = v|_{\mathcal{Y}} \in \mathcal{N}'$. Therefore, the statement follows from induction. □

Let $\underline{\mathcal{X}} = \{x \in \mathcal{X} : v(x) \neq 0 \text{ for some } v \in \mathcal{N}\}$. Define a relation $\sim$ on $\underline{\mathcal{X}}$ via

$$x \sim y \quad\Longleftrightarrow\quad v(y) \neq 0 \text{ for all } v \in \mathcal{N} \text{ such that } v(x) \neq 0.$$

It is easy to see that $\sim$ is an equivalence relation: If there exist $v, w \in \mathcal{N}$ such that $v(y) \neq 0 = v(x)$ and $w(x) \neq 0 \neq w(y)$, then $u := v(y)w - w(y)v \in \mathcal{N}$ satisfies $u(y) = 0 \neq u(x)$, and so $\sim$ is symmetric. Transitivity can be shown similarly. In the language of matroid theory the equivalence classes are the coparallel classes.

Lemma 4.5. A subset $\mathcal{Z} \subseteq \mathcal{X}$ is an equivalence class of $\sim$ if and only if there exist circuits $\sigma_0, \sigma_1, \ldots, \sigma_s$ of $\mathcal{N}$ such that

$$\mathcal{Z} = \sigma_0 \setminus \cup_{j=1}^{s}\sigma_j,$$

and such that $\mathcal{Z} \setminus \sigma \in \{\emptyset, \mathcal{Z}\}$ for all circuits σ of $\mathcal{N}$.

P r o o f. If $x \not\sim y$ for some $y \in \mathcal{X}$, then there exists a $v \in \mathcal{N}$ such that $v(x) \neq 0$ and $v(y) = 0$. By Lemma 2.2 there exists a circuit with the same property. Conversely, if $y \sim x$, then $y \in \sigma$ for any circuit σ such that $x \in \sigma$. □

Let $C \in \mathbb{R}^{c\times\mathcal{X}}$ be a matrix such that the rows $c_1, \ldots, c_c$ of C form a circuit basis of $\mathcal{N}$. Since each circuit basis contains a basis, the rank of C equals the dimension of $\mathcal{N}$. The columns of C are denoted by $\{C_x\}_{x\in\mathcal{X}}$.

Lemma 4.6. Let $\mathcal{Z}$ be an equivalence class of $\sim$. The rank of the submatrix $C|_{\mathcal{Z}}$ consisting of those columns C_x indexed by $\mathcal{Z}$ is one.

P r o o f. Let $\mathcal{Z} \subseteq \mathcal{X}$. If the rank of $C|_{\mathcal{Z}}$ is larger than one, then there exist two circuit vectors c_1, c_2 such that $c_1|_{\mathcal{Z}}$ and $c_2|_{\mathcal{Z}}$ are linearly independent and have support $\mathcal{Z}$. Let $x \in \mathcal{Z}$. Let $v = c_2(x)c_1 - c_1(x)c_2 \in \mathcal{N}$. Then $v|_{\mathcal{Z}} \neq 0$ and $\operatorname{supp}(v|_{\mathcal{Z}}) \subseteq \mathcal{Z} \setminus \{x\}$. Therefore, $\mathcal{Z}$ is not an equivalence class of $\sim$. □

The main argument of the last proof can be reformulated in terms of the elimination axiom of oriented matroid theory, cf. [17]. In the language of matroid theory Lemma 4.6 states that the coparallel classes of a matroid have corank one.

P r o o f o f T h e o r e m 4.1. Suppose $\max D_{\mathcal{E}} = \log(2)$. By Lemma 4.6, the rank of C is bounded from above by the number of equivalence classes of $\sim$. Let $\mathcal{Z}$ be an equivalence class of $\sim$. By definition, the submatrix $C|_{\mathcal{Z}} \in \mathbb{R}^{c\times\mathcal{Z}}$ is not the zero matrix.

By Lemmas 4.4 and 4.5 the rows $c_i|_{\mathcal{Z}}$ of $C|_{\mathcal{Z}}$ satisfy $\sum_{x\in\mathcal{Z}} c_i(x) = 0$. Hence each equivalence class must contain at least two elements. Therefore, the rank of C, which equals the codimension of $\mathcal{E}$, is at most $\lfloor\frac{N}{2}\rfloor$, and so the dimension of $\mathcal{E}$ is bounded from below by $N - 1 - \lfloor\frac{N}{2}\rfloor = \lceil\frac{N}{2}\rceil - 1$. $\square$

Lemma 4.7. If the dimension of $\mathcal{N}$ equals the number of equivalence classes of $\sim$, then the equivalence classes are the circuits of $\mathcal{N}$. In other words, the circuit vectors $c_1, \dots, c_c$ of a circuit basis are in bijection with the equivalence classes $\mathcal{Z}_1, \dots, \mathcal{Z}_c$, such that $\mathcal{Z}_i = \operatorname{supp}(c_i)$. Hence $\overline{\mathcal{E}}$ is the mixture of $\overline{\mathcal{E}}_1, \dots, \overline{\mathcal{E}_c}$, where $\mathcal{E}_c$ is the exponential family $\overline{\mathcal{E}} \cap \mathbf{P}(\mathcal{Z}_i)^\circ$.

Proof. Let $\mathcal{Z}_1, \dots, \mathcal{Z}_{c'}$ be the set of equivalence classes of $\sim$. Reorder $\mathcal{X}$ such that the equivalence classes are given by consecutive numbers. Let $\tilde{C}$ be the matrix obtained from C by doing a Gauss elimination through row operations. By assumption $\tilde{C}$ has $\dim\mathcal{N} = c'$ nonzero rows. By Lemma 4.6, the ith row $\tilde{c}_i$ of $\tilde{C}$ has support contained in $\mathcal{Z}_i \cup \dots \cup \mathcal{Z}_{c'}$. In particular, $\operatorname{supp}(\tilde{c}_{c'}) = \mathcal{Z}_{c'}$. Therefore, $\tilde{c}_{c'}$ is a circuit vector. If $v \in \mathcal{N}$ has $v(x) \neq 0$ for some $x \in \mathcal{Z}_{c'}$, then $\tilde{v} = v - \frac{v(x)}{\tilde{c}_{c'}(x)}\tilde{c}_{c'}(x)$ satisfies $\operatorname{supp}(\tilde{v}) = \operatorname{supp}(v)\setminus\mathcal{Z}_{c'}$. Hence no other circuit intersects $\mathcal{Z}_{c'}$. By induction, $\operatorname{supp}(c_i)$ equals an equivalence class of $\sim$ for each i. The first statement follows from $\operatorname{supp}(c_i) \neq \operatorname{supp}(c_j)$ for $1 \leq i < j \leq c$. The last statement is a consequence of Corollary 2.4. $\square$

Proof of Theorem 4.2. Assume that the dimension of $\mathcal{E}$ equals $\lceil\frac{N}{2}\rceil - 1$. By the proof of Theorem 4.1 there must be $m := \lfloor\frac{N}{2}\rfloor$ equivalence classes of $\sim$. If N is even, then each equivalence class has cardinality two. If N is odd, then there may be one equivalence class $\mathcal{Z}$ of cardinality three. In this case, reorder $\mathcal{X}$ such that $\mathcal{Z} = \{N-2, N-1, N\}$. By Lemma 4.7 there exists a circuit vector $c \in \mathcal{N}$ such that $\operatorname{supp}(c) = \mathcal{Z}$. Assume without loss of generality that c_{N-2} and c_{N-1} are positive and that $c_N = -(c_{N-1} + c_{N-2}) = -1$. Then

$$\sum_{i=N-2}^{N} c_i \log|c_i| = -h(c_{N-1}, c_{N-2}) \neq 0,$$

where $h(p, q)$ is the entropy of a binary random variable with probabilities p, q. Therefore, if N is even or if $\mathbf{1}$ is a reference measure of $\mathcal{E}$, then all equivalence classes of $\sim$ have cardinality two.

If $\underline{\mathcal{X}} = \mathcal{X}$, then, by Lemma 4.7, there are exponential families $\mathcal{E}_1, \dots, \mathcal{E}_c$ such that $\mathcal{E}_i \subseteq \mathbf{P}(\mathcal{Z}_i)^\circ$ for $i = 1, \dots, c$ and such that $\overline{\mathcal{E}}$ is the mixture of $\overline{\mathcal{E}_1}, \dots, \overline{\mathcal{E}_c}$. For $i = 1, \dots, c$ there is a unique circuit vector with support $\mathcal{Z}_i$, hence $\mathcal{E}_i \neq \mathbf{P}(\mathcal{Z}_i)^\circ$, so $\mathcal{E}_i$ has dimension $|\mathcal{Z}_i| - 1$. If $|\mathcal{Z}_i| = 2$, then $\mathcal{E}_i$ consists of the uniform distribution $\frac{1}{2}\mathbf{1}_{\mathcal{Z}_i}$ on $\mathcal{Z}_i$, and so the mixture of $\overline{\mathcal{E}_i}$ for those i satisfying $|\mathcal{Z}_i| = 2$ is a partition model. If $\underline{\mathcal{X}} \neq \mathcal{X}$, then $\mathcal{X} \setminus \underline{\mathcal{X}}$ is a singleton, say $\{N\}$, and by Lemma 4.7, $\overline{\mathcal{E}}$ equals the mixture of the point measure δ_N and $\overline{\mathcal{E}} \cap \mathbf{P}(\underline{\mathcal{X}})$. The same analysis as above applies to $\overline{\mathcal{E}} \cap \mathbf{P}(\underline{\mathcal{X}})$, and one can choose $\overline{\mathcal{E}}_{\mathcal{Z}}$ to be the mixture of δ_N and $\overline{\mathcal{E}_c}$. $\square$

Proof of Proposition 4.3. Let $\mathcal{E}$ be a one-dimensional exponential family with normal space $\mathbb{R}u$. Without loss of generality assume that u^+ and u^- are probability measures. By Corollary 2.9 the set of local maximizers of $D_{\mathcal{E}}$ consists of u^+ and u^-. $\mathcal{E}$ satisfies $\max D_{\mathcal{E}} = \log(2)$ if and only if $(u^+)_{\mathcal{E}} = (u^-)_{\mathcal{E}} = \frac{1}{2}(u^+ + u^-)$, which happens if and only if $u^+ + u^-$ is a reference measure of $\mathcal{E}$, proving existence and uniqueness of $\mathcal{E}$. □

5. OPTIMAL EXPONENTIAL FAMILIES

Corollary 2.10 says that $\max D_{\mathcal{E}} \geq \log(2)$ for all exponential families $\mathcal{E} \neq \mathbf{P}(\mathcal{X})^{\circ}$. Therefore D-optimality is only interesting for $D \geq \log(2)$. The case $D = \log(2)$ was studied in Section 4, where it was shown that $D_{N,k} = \log(2)$ if and only if $\lceil \frac{N}{2} \rceil - 1 \leq k < N$. This condition is equivalent to $\lceil \frac{N}{k+1} \rceil = 2$. Many $\log(2)$-dimension optimal exponential families are partition exponential families.

Example 5.1. Any zero-dimensional exponential family $\mathcal{E} = \{\nu\}$ is dimension-optimal. The function $P \mapsto D(P\|\nu)$ is convex on the probability simplex $\mathbf{P}(\mathcal{X})$ and attains its maximum at a vertex of $\mathbf{P}(\mathcal{X})$, which corresponds to a point distribution. Therefore,

$$\max D_{\mathcal{E}} = \max\{-\log(\nu_x) : x \in \mathcal{X}\} \geq \log|\mathcal{X}|.$$

Hence $D_{N,1} = \log(N)$, and $\mathcal{E}$ is D-optimal if and only if $\nu_x \geq e^{-D}$ for all $x \in \mathcal{X}$. Zero-dimensional exponential families are the dimension D-optimal exponential families for $D \geq \log|\mathcal{X}|$. In general, they are not the only inclusion D-optimal exponential families, see Example 5.2.

Example 5.2. Let $\mathcal{X} = \{1, 2, 3\}$. Any zero-dimensional exponential family $\mathcal{E} = \{\nu\}$ satisfies $\max D_{\mathcal{E}} \geq \log(3)$. Therefore, if $\log(2) \leq D < \log(3)$, then the dimension D-optimal exponential families are one-dimensional. The normal space $\mathcal{N}$ of any one-dimensional exponential family $\mathcal{E}$ is spanned by a single element u, which can be taken to be normalized, such that $\partial\mathbf{U}_{\mathcal{N}} = \{\pm u\}$. By Corollary 2.9 the set of local maximizers of $D_{\mathcal{E}}$ equals $\{u^+, u^-\}$. Let $P_{\mathcal{E}} = (u^+)_{\mathcal{E}} = (u^-)_{\mathcal{E}}$, then $P_{\mathcal{E}} = \mu u^+ + (1-\mu)u^-$ for some $0 < \mu < 1$. Hence $D_{\mathcal{E}}(u^+) = -\log\mu$ and $D_{\mathcal{E}}(u^-) = -\log(1-\mu)$. It follows that $\mathcal{E}$ is dimension D-optimal if and only $e^{-D} \leq \mu \leq 1 - e^{-D}$. Alternatively, using Theorem 2.8, $\mathcal{E}$ is dimension D-optimal if and only if $-\log(e^D - 1) \leq \overline{D}_{\mathcal{E}}(u) \leq \log(e^D - 1)$.

If $D \geq \log(3)$, then the dimension D-optimal exponential families are zero-dimensional, consisting of a single point $\{\nu\}$ such that $\min\{\nu_1, \nu_2, \nu_3\} \geq e^{-D}$. There are also one-dimensional inclusion D-optimal exponential families: Consider, for example, the exponential family $\mathcal{E}$ with sufficient statistics $A = (0, 1, 2)$ and reference measure $\nu = (1, 4, 1)$ (see Fig. 1). The two local maximizers are $u^+ = \delta_2$ and $u^- = \frac{1}{2}(\delta_1 + \delta_3)$. Their rI-projection is $P_{\mathcal{E}} = \frac{1}{6}\nu$. Hence $D_{\mathcal{E}}(u^+) = \log\frac{3}{2}$ and $D_{\mathcal{E}}(u^-) = \log 3$, and so $\max D_{\mathcal{E}} = \log 3$. By Theorem 2.3, $\mathcal{E}$ does not contain the uniform distribution. Therefore, any point $P \in \mathcal{E}$ satisfies $\max D(\cdot\|P) > \max D_{\mathcal{E}}$.

The following theorem generalizes the special case of Theorem 4.2 when N is even.

Theorem 5.3. Let $\mathcal{X}$ be a finite set of cardinality N. Then $D_{N,k} \geq \log(N/(k+1))$ for all $0 \leq k < N$. If $\mathcal{E}$ is a k-dimensional exponential family that satisfies $\max D_{\mathcal{E}} =$

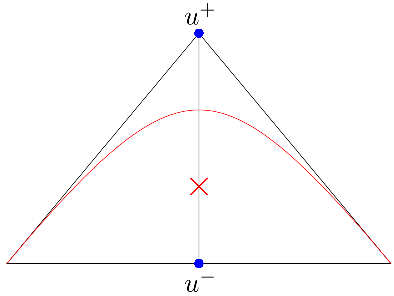

Fig. 1. The exponential family with $A = (0, 1, 2)$ and $\nu = (1, 4, 1)$ from Example 5.2. The vertical line equals the shift of the normal space passing through the two local maximizers u^+ and u^-. The cross marks the uniform distribution.

$\log(N/(k+1))$, then $\mathcal{E}$ is a partition model of a homogeneous partition of coarseness $N/(k+1)$. In particular, if N is divisible by $(k+1)$, then $D_{N,k} = \log(N/(k+1))$, and the dimension $D_{N,k}$-optimal models are partition models.

Proof. First assume that $\mathcal{E} \in \mathcal{H}_{\mathbf{1}}$. Let A be a sufficient statistics of $\mathcal{E}$. The moment map π_A maps the uniform distribution $Q = \frac{1}{N}\mathbf{1}$ to a point in the relative interior of the convex support $\mathbf{M}_A$ (as defined in Section 2.2). By Carathéodory's theorem there are $k+1$ vertices $A_{x_0}, \dots, A_{x_k}$ of $\mathbf{M}_A$ and non-negative real numbers $\lambda_0, \dots, \lambda_k$ such that $\pi_A(Q) = \sum_{i=0}^k \lambda_i A_{x_i}$ and $\sum_{i=0}^k \lambda_i = 1$. Let $P = \sum_{i=0}^k \lambda_i \delta_{x_i}$, then $Q = P_{\mathcal{E}}$. By the Pythagorean theorem, $\max D_{\mathcal{E}} \geq D_{\mathcal{E}}(P) = H(Q) - H(P) \geq \log(N) - \log(k+1)$, proving the first assertion.

If equality holds, then $H(P) = \log(k+1)$, and so $\lambda_0 = \dots = \lambda_k = \frac{1}{k+1}$. Let $x \in \mathcal{X} \setminus \{x_0, \dots, x_k\}$. For $i \in \{0, \dots, k\}$ let C_i be the convex hull of $\{A_x\} \cup \{A_{x_0}, \dots, A_{x_k}\} \setminus \{A_{x_i}\}$. By Carathéodory's theorem the sets C_i cover the convex hull of $A_{x_0}, \dots, A_{x_k}$ and A_x. In particular, $\pi_A(Q) \in C_j$ for some $j \in \{0, \dots, k\}$, so $\pi_A(Q) = \sum_{i \neq j} \lambda'_i A_{x_i} + \lambda'_j A_x$, where $\lambda'_i \geq 0$ and $\sum_{i=0}^k \lambda_i = 1$. By the same argument as above it follows that $\lambda'_0 = \dots = \lambda'_k = \frac{1}{k+1}$; for otherwise the probability measure $P' = \sum_{i \neq j} \lambda'_i \delta_{x_i} + \lambda'_0 \delta_x$ would satisfy $D_{\mathcal{E}}(P') > \log(N/(k+1))$. Therefore, $A_x = (k+1)\pi_A(Q) - \sum_{i \neq j} A_{x_i} = A_{x_j}$. Hence each column A_x of A is equal to one of the $k+1$ columns $A_{x_0}, \dots, A_{x_k}$.

Let $\sim$ be the equivalence relation on $\mathcal{X}$ defined by $x \sim y$ if and only if $A_x = A_y$, and let $\mathcal{X}' = (\mathcal{X}^1, \dots, \mathcal{X}^{N'})$ be the corresponding partition into equivalence classes. Then $N' \leq k+1$ by what was shown until now. From $\dim(\mathcal{E}) = \dim(\mathbf{M}_A)$ one concludes $N' = k+1$, and $\mathbf{M}_A$ is a simplex of dimension k. By Lemma 3.4, $\mathcal{E}$ equals the partition model of $\mathcal{X}'$. Lemma 3.5 implies that the coarseness of $\mathcal{X}'$ equals $\frac{N}{k+1}$, which must be an integer. Furthermore, $\mathcal{X}'$ is homogeneous.

It remains to prove $\max D_{\mathcal{E}} > \log(N/(k+1))$ in the case $\mathcal{E} \notin \mathcal{H}_{\mathbf{1}}$. Let $Q_{\mathcal{E}}$ be the rI-projection of the uniform distribution Q, and let $\mathcal{N}_{\mathbf{1}}$ be the set of probability distributions that rI-project to $Q_{\mathcal{E}}$. Equivalently, $\mathcal{N}_{\mathbf{1}} = (Q + \mathcal{N}) \cap \mathbf{P}(\mathcal{X})$. On the convex set $\mathcal{N}_{\mathbf{1}}$, the function $D_{\mathcal{E}}(P) = D(P\|\nu) - D(P_{\mathcal{E}}\|\nu) = D(P\|\nu) - D(Q_{\mathcal{E}}\|\nu)$ is convex, since the information divergence is convex in its first argument. Hence $D_{\mathcal{E}}$ is maximal at the vertices of $\mathcal{N}_{\mathbf{1}}$. Let P be a vertex of $\mathcal{N}_{\mathbf{1}}$. Assume that $v \in \mathcal{N}$ satisfies $\mathrm{supp}(v) \subseteq \mathrm{supp}(P)$. Then there exists $\epsilon > 0$ such that $P \pm \epsilon v \in \mathcal{N}_{\mathbf{1}}$ and $P = \frac{1}{2}(P+\epsilon v) + \frac{1}{2}(P-\epsilon v)$. Hence $v = 0$. Therefore, the set $\{A_x : P(x) > 0\}$ is linearly independent. In particular $|\mathrm{supp}(P)| \leq \dim(\mathcal{X}) + 1$.

Denote by $\mathcal{E}_{\mathbf{1}}$ the exponential family with uniform reference measure and with the same normal space as $\mathcal{E}$. On $\mathcal{N}_{\mathbf{1}}$ the difference

$$\delta(P) := D_{\mathcal{E}}(P) - D_{\mathcal{E}_{\mathbf{1}}}(P) = -\sum_{x\in\mathcal{X}} P(x)\log P_{\mathcal{E}}(x) - \log N$$

is affine and positive at the uniform distribution. So there is a vertex P of $\mathcal{N}_{\mathbf{1}}$ with $\delta(P) > 0$, and so $D_{\mathcal{E}}(P) > D_{\mathcal{E}_{\mathbf{1}}}(P) = \log N - H(P) \geq \log(N/(k+1))$. $\square$

The value of $D_{N,k}$ is unknown when $k+1$ does not divide N. The situation is known for $N = 3$, see Example 5.2: If $1 \leq k < 3$, then $D_{N,k} = \log(2)$, and all dimension $D_{N,1}$-optimal exponential families that contain the uniform distribution are partition models. The following conjecture generalizes this example and Theorems 4.2 and 5.3:

Conjecture 5.4. $D_{N,k} = \log\lceil\frac{N}{k+1}\rceil$, and the dimension $D_{N,k}$-optimal exponential families containing the uniform distribution are partition models.

The following weaker statement holds:

Lemma 5.5. Let $\mathcal{X}' = \left\{\mathcal{X}^1, \dots, \mathcal{X}^{N'}\right\}$ be a partition of coarseness $c < N$ such that $\mathcal{X}^1$ has cardinality $l \leq c$ and all other components $\mathcal{X}^i$ for $i > 1$ have cardinality c. Then the partition model $\mathcal{E}$ of $\mathcal{X}'$ is $\log(c)$-inclusion optimal.

Proof. The fact that $\max D_{\mathcal{E}} = \log(c)$ follows from Lemma 3.5. It remains to prove the optimality. Let $\mathcal{E}' \subseteq \mathcal{E}$ be an exponential family contained in $\mathcal{E}$. Let $\mathcal{Z}$ be the union of all blocks of $\mathcal{X}'$ of cardinality c. Assume that there exists a probability measure $Q \in \overline{\mathcal{E}} \setminus \overline{\mathcal{E}'}$ with support contained in $\mathcal{Z}$. By Corollary 3.6 there exists $P \in \mathbf{P}(\mathcal{Z})$ such that $Q = P_{\mathcal{E}}$ and $D(P\|Q) = \log(c)$. Let $Q' = P_{\mathcal{E}'} \in \mathcal{E}$. Then $D(P\|Q') = D(P\|Q) + D(Q\|Q') > \log(c)$ by the Pythagorean identity. Otherwise, if $\overline{\mathcal{E}} \cap \mathbf{P}(\mathcal{Z}) = \overline{\mathcal{E}'} \cap \mathbf{P}(\mathcal{Z})$, then $\dim(\mathcal{E}) = \dim(\overline{\mathcal{E}} \cap \mathbf{P}(\mathcal{Y})) + 1 = \dim(\overline{\mathcal{E}}' \cap \mathbf{P}(\mathcal{Y})) + 1 \leq \dim(\mathcal{E}')$, so $\mathcal{E} = \mathcal{E}'$. $\square$

Theorem 5.3 can be applied to the hierarchical models $\overline{\mathcal{E}}_K$ for $K \subseteq [n]$ introduced in Remark 3.7. By Theorem 5.3 the hierarchical model $\overline{\mathcal{E}}_K$ is dimension optimal with $\max D(\cdot\|\mathcal{E}_K) = \sum_{i\in[n]\setminus K}\log(N_i)$. If $N_n = 2$, then the choice $K = \{1, \dots, n-1\}$ yields an exponential family of dimension less than $|\mathcal{X}|/2$ such that $\max D(\cdot\|\mathcal{E}_K) = \log(2)$, and Theorem 4.1 implies that $\mathcal{E}_K$ is dimension optimal. The following proposition says that the exponential families $\mathcal{E}_K$ are the unique dimension D-optimal hierarchical models for many values of D.

Proposition 5.6. Let $\mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_n$, where $N_i = |\mathcal{X}_i| < \infty$. For any $K \subseteq [n]$ let $D_K = \sum_{i\notin K}\log(N_i)$. The hierarchical model $\mathcal{E}_K$ is dimension D_K-optimal.

Let l be any divisor of $N := |\mathcal{X}| = \prod_{i=1}^n N_i$. If $\mathcal{E}$ is any hierarchical model that is dimension $\log(N/l)$-optimal, then there is a subset $K \subseteq [n]$ such that $\mathcal{E} = \mathcal{E}_K$.

The proposition implies that if l is not of the form $\prod_{i\in K} N_i$ for some subset $K \subseteq [n]$, then there is no hierarchical model that is dimension $\log(N/l)$-optimal.

Proof. It only remains to prove the last statement. If $\mathcal{E}$ satisfies the assumptions, then $\mathcal{E}$ is a partition model by Theorem 5.3. Therefore, it suffices to prove that any hierarchical model that is also a partition model is of the form $\overline{\mathcal{E}_K}$.

Let Δ be a simplicial complex on $[n]$ such that $\mathcal{E} = \mathcal{E}_\Delta$, and let $K = \cup_{J\in\Delta} J$. Then $\mathcal{E}$ is a submodel of $\mathcal{E}_K$. Let A be a sufficient statistics of $\mathcal{E}$. By Lemma 2.6 the convex supports of $\mathcal{E}$ and $\mathcal{E}_K$ have the same number of vertices. By Lemma 3.4 both are simplices, hence they have the same dimension, so $\mathcal{E} = \mathcal{E}_K$. □

6. DISCUSSION

Conjecture 5.4 would imply that the partition models of Lemma 5.5 are dimension optimal among all exponential families. If the conjecture were true, then it would suggest the following interpretation: In many cases the information divergence $D(P\|Q)$ can be interpreted as the information which is lost when P is the true probability distribution, but computations are carried out with Q. For example, in the case of the independence model $\mathcal{E}_1$ of two variables, $D_{\mathcal{E}_1}$ equals the mutual information and measures the amount of information that one variable carries about the other variable. If a probability measure is replaced by its rI-projection, then this information is lost.

For the exponential families $\mathcal{E}_K$ the loss equals $D_K = \sum_{i\notin K} \log(N_i)$, which is precisely the maximal information that the random variables that are not in K can carry. Assuming that the conjecture is true, if the model is smaller than $\mathcal{E}_K$, then, in general, more information can be lost. In this interpretation the fact that $\max D_{\mathcal{E}} \geq \log(2)$ unless $\mathcal{E} = \mathbf{P}(\mathcal{X})^\circ$ means that for any exponential family $\mathcal{E} \neq \mathbf{P}(\mathcal{X})^\circ$ in general at least one bit is necessary to compensate the approximation of arbitrary probability measures.

ACKNOWLEDGEMENT

I thank Yaroslav Bulatov, who first asked the main question studied in this work. Further thanks goes to Nihat Ay, whose questions led me in similar directions, but from a different starting point.

(Received June 28, 2012)

REFERENCES

[1] N. Ay: An information-geometric approach to a theory of pragmatic structuring. Ann. Probab. *30* (2002), 416–436.

[2] N. Ay: Locality of global stochastic interaction in directed acyclic networks. Neural Computat. *14* (2002), 2959–2980.

[3] L. Brown: Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory. Institute of Mathematical Statistics, Hayworth 1986.

[4] T. Cover and J. Thomas: Elements of Information Theory. First edition. Wiley, 1991.

[5] I. Csiszár and P. Shields: Information Theory and Statistics: A Tutorial. First edition. Foundations and Trends in Communications and Information Theory. Now Publishers, 2004.

[6] I. Csiszár and F. Matúš: Generalized maximum likelihood extimates for exponential families. Probab. Theory Rel. Fields *141* (2008), 213–246.

[7] S. Della Pietra, V. Della Pietra, and J. Lafferty: Inducing features of random fields. IEEE Trans. Pattern Analysis and Machine Intelligence *19* (1997), 380–393.

[8] M. Drton, B. Sturmfels, and S. Sullivant: Lectures on algebraic statistics. In: Oberwolfach Seminars *39*, Birkhäuser, Basel 2009.

[9] D. Geiger, C. Meek, and B. Sturmfels: On the toric algebra of graphical models. Ann. Statist. *34* (2006), 5, 1463–1492.

[10] E. T. Jaynes: Information theory and statistical mechanics. Phys. Rev. *106* (1957), 4, 620–630.

[11] J. Juríček: Maximization of information divergence from multinomial distributions. Acta Univ. Carolin. *52* (2011), 1, 27–35.

[12] S. L. Lauritzen: Graphical Models. First edition. Oxford Statistical Science Series, Oxford University Press, 1996.

[13] R. Linsker: Self-organization in a perceptual network. IEEE Computer *21* (1988), 105–117.

[14] F. Matúš and N. Ay: On maximization of the information divergence from an exponential family. In: Proc. WUPES'03, University of Economics, Prague 2003, pp. 199–204.

[15] F. Matúš and J. Rauh: Maximization of the information divergence from an exponential family and criticality. In: 2011 IEEE International Symposium on Information Theory Proceedings (ISIT2011), 2011.

[16] G. Montúfar, J. Rauh, and N. Ay: Expressive power and approximation errors of Restricted Boltzmann Machines. In: NIPS, 2011.

[17] J. Oxley: Matroid Theory. First edition. Oxford University Press, New York 1992.

[18] J. Rauh: Finding the Maximizers of the Information Divergence from an Exponential Family. Ph.D. Dissertation, Universität Leipzig, 2011.

[19] J. Rauh: Finding the maximizers of the information divergence from an exponential family. IEEE Trans. Inform. Theory *57* (2011), 6, 3236–3247.

[20] J. Rauh, T. Kahle, and N. Ay: Support sets of exponential families and oriented matroids. Internat. J. Approx. Reasoning *52* (2011), 5, 613–626.

[21] S. C. Zhu, Y. N. Wu, and D. Mumford: Minimax entropy principle and its application to texture modeling. Neural Computation *9* (1997), 1627–1660.

Johannes Rauh, MPI MIS, Inselstraße 23, 04103 Leipzig. Germany.
e-mail: jrauh@mis.mpg.de