KYBERNETIKA — VOLUME 42 (2006), NUMBER 5, PAGES 517 – 538

MAXIMIZING MULTI–INFORMATION

Nihat Ay and Andreas Knauf

Stochastic interdependence of a probability distribution on a product space is measured by its Kullback–Leibler distance from the exponential family of product distributions (called multi-information). Here we investigate low-dimensional exponential families that contain the maximizers of stochastic interdependence in their closure.

Based on a detailed description of the structure of probability distributions with globally maximal multi-information we obtain our main result: The exponential family of pure pair-interactions contains all global maximizers of the multi-information in its closure.

Keywords: multi-information, exponential family, relative entropy, pair-interaction, infomax principle, Boltzmann machine, neural networks

AMS Subject Classification: 82C32, 92B20, 94A15

1. INTRODUCTION

The starting point of this article is a geometric interpretation of the interdependence[1] of stochastic units. In order to illustrate the basic idea, we consider two units with the configuration sets $\Omega_1 = \Omega_2 = \{0, 1\}$. The configuration set of the whole system is just the Cartesian product $\Omega_1 \times \Omega_2 = \{(0,0),(1,0),(0,1),(1,1)\}$. The set of probability distributions (*states*) is a three-dimensional simplex $\overline{\mathcal{P}}(\Omega_1 \times \Omega_2)$ with the four extreme points $\delta_{(\omega_1,\omega_2)}$, $\omega_1, \omega_2 \in \{0,1\}$ (Dirac measures). The two units are independent with respect to $p \in \overline{\mathcal{P}}(\Omega_1 \times \Omega_2)$ iff

$$p(\omega_1, \omega_2) \;=\; p_1(\omega_1)\, p_2(\omega_2) \quad \text{for all} \quad (\omega_1, \omega_2) \in \Omega_1 \times \Omega_2. \tag{1.1}$$

The set of *factorizable* distributions (1.1) is a two-dimensional manifold $\mathcal{F}$. Figure 1 shows the simplex $\overline{\mathcal{P}}(\Omega_1 \times \Omega_2)$ and its submanifold $\mathcal{F}$.

Given an arbitrary probability distribution p, we quantify the interdependence of the two units with respect to p by its Kullback–Leibler distance from the set $\mathcal{F}$. In our two-unit case, this distance is nothing but the well known mutual information, which has been introduced by Shannon [10] as a fundamental quantity that provides a measure of the capacity of a communication channel.

[1] Throughout the paper we use the term *interdependence* to indicate stochastic dependence among units, as opposed to dependence of general random variables.

Motivated by so-called *Infomax principles* within the field of neural networks [8, 11], one of us has investigated maximizers of the interdependence [6, 7] of stochastic units. In our two-unit example, these are the distributions

$$\frac{1}{2}\left(\delta_{(0,0)}+\delta_{(1,1)}\right), \quad \text{and} \quad \frac{1}{2}\left(\delta_{(1,0)}+\delta_{(0,1)}\right) \qquad \text{(see Figure 1).}$$

This article continues that work by analyzing the structure of maximizers of stochastic interdependence. In particular, this leads to some answers to the question on the existence and the structure of a natural low dimensional manifold that contains all maximizers of the stochastic interdependence (see [6], 3.4 (ii) and [7], 4.2.3). We will prove that the exponential family of pure pair-interactions contains the global maximizers of multi-information in its closure. In our example of two binary units this exponential family is given by the convex hull of the two maximizers $\frac{1}{2}\left(\delta_{(0,0)}+\delta_{(1,1)}\right)$ and $\frac{1}{2}\left(\delta_{(1,0)}+\delta_{(0,1)}\right)$ shown in Figure 1.

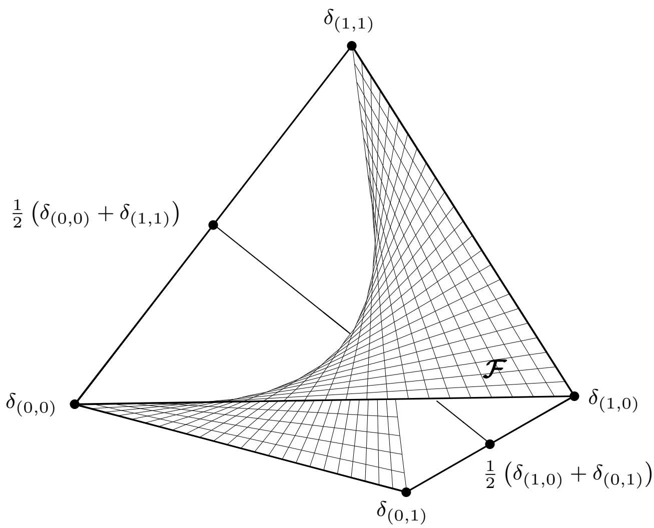

Fig. 1. The exponential family $\mathcal{F}$ in the simplex of probability distributions.

In physics, pair interactions are considered as fundamental mechanisms that underly most theories. Within the field of neural networks, the physical concept of pair-interactions is used to model the synaptic interactions of neurons.

2. NOTATION

Let Ω be a nonempty and finite set. In the corresponding real vector space $\mathbb{R}^\Omega$, we have the canonical basis e_ω, $\omega \in \Omega$, which induces the natural scalar product $\langle \cdot, \cdot \rangle$. The set of probability distributions on Ω is denoted by $\overline{\mathcal{P}}(\Omega)$:

$$\overline{\mathcal{P}}(\Omega) := \left\{ p = \left(p(\omega)\right)_{\omega\in\Omega} \in \mathbb{R}^\Omega : p(\omega) \geq 0 \text{ for all } \omega, \text{ and } \textstyle\sum_{\omega\in\Omega} p(\omega) = 1 \right\}.$$

For a probability distribution p, we consider its support $\operatorname{supp} p := \{\omega \in \Omega : p(\omega) > 0\}$. The strictly positive distributions $\mathcal{P}(\Omega)$ have maximal support Ω:

$$\mathcal{P}(\Omega) := \{p \in \overline{\mathcal{P}}(\Omega) : \operatorname{supp} p = \Omega\}.$$

Note that $\overline{\mathcal{P}}(\Omega)$ is the closure of $\mathcal{P}(\Omega)$. For every vector $X = (X(\omega))_{\omega \in \Omega} \in \mathbb{R}^\Omega$, we consider the corresponding *Gibbs measure*:

$$\exp(X) \in \mathcal{P}(\Omega), \quad \exp(X)(\omega) := \frac{e^{X(\omega)}}{\sum_{\omega' \in \Omega} e^{X(\omega')}}.$$

The image $\exp(\mathcal{T})$ of a linear (or more generally affine) subspace $\mathcal{T}$ of $\mathbb{R}^\Omega$ with respect to the map $X \mapsto \exp(X)$ is called *exponential family* (*induced by* $\mathcal{T}$).

In this article, we are mainly interested in the "distance" of probability distributions from a given exponential family $\mathcal{E}$. More precisely, we use the *Kullback–Leibler divergence* or *relative entropy* $D : \overline{\mathcal{P}}(\Omega) \times \overline{\mathcal{P}}(\Omega) \to [0, \infty) \cup \{\infty\}$,

$$(p, q) \mapsto D(p \,\|\, q) := \begin{cases} \sum_{\omega \in \operatorname{supp} p} p(\omega) \ln \frac{p(\omega)}{q(\omega)}, & \text{if } \operatorname{supp} p \subset \operatorname{supp} q, \\ \infty, & \text{otherwise,} \end{cases}$$

to define the continuous[2] function $D_{\mathcal{E}} : \overline{\mathcal{P}}(\Omega) \to \mathbb{R}_+$,

$$p \mapsto D_{\mathcal{E}}(p) := \inf_{q \in \mathcal{E}} D(p \,\|\, q).$$

For $k \in \mathbb{N}$ we denote the set $\{1, \ldots, k\}$ by $[k]$.

3. SUFFICIENCY OF LOW–DIMENSIONAL EXPONENTIAL FAMILIES FOR THE MAXIMIZATION OF MULTI–INFORMATION

We consider the set $V := [N] = \{1, \ldots, N\}$ of $N \geq 2$ *units*, and corresponding sets Ω_i, $i \in [N]$, of *configurations*. The number $|\Omega_i|$ of configurations of a unit i is denoted by n_i. Without restriction of generality we assume

$$2 \leq n_1 \leq n_2 \leq \cdots \leq n_N.$$

For a subsystem $A \subseteq [N]$, the set of configurations on A is given by the product $\Omega_A := \times_{i \in A} \Omega_i$. One has the natural restriction

$$X_A : \Omega_V \to \Omega_A, \qquad (\omega_i)_{i \in [N]} \mapsto (\omega_i)_{i \in A},$$

which induces the projection

$$\overline{\mathcal{P}}(\Omega_V) \to \overline{\mathcal{P}}(\Omega_A), \qquad p \mapsto p_A,$$

where p_A denotes the image measure of p with respect to the variable X_A. For $i \in [N]$ we write p_i instead of $p_{\{i\}}$.

[2] See Lemma 4.2 of [6] for a proof.

A probability distribution $p \in \overline{\mathcal{P}}(\Omega_V)$ is called *factorizable* if it satisfies

$$p(\omega_1, \ldots, \omega_N) = p_1(\omega_1) \cdot \ldots \cdot p_N(\omega_N) \quad \text{for all} \quad (\omega_1, \ldots, \omega_N) \in \Omega_V.$$

The set $\mathcal{F}$ of strictly positive and factorizable probability distributions on Ω_V is an exponential family in $\mathcal{P}(\Omega_V)$ with

$$\dim \mathcal{F} = \sum_{i=1}^{N} (n_i - 1).$$

Now let us consider the function $D_{\mathcal{F}}$, which measures the distance from $\mathcal{F}$. We have $D_{\mathcal{F}}(p) = 0$ if and only if $p \in \overline{\mathcal{P}}(\Omega_V)$ is factorizable. Thus, this distance function can be interpreted as a measure that quantifies the stochastic interdependence of the units in $[N]$. The following entropic representation of $D_{\mathcal{F}}$ is well known (see [4]):

$$I_p(X_1, \ldots, X_N) := D_{\mathcal{F}}(p) = \sum_{i=1}^{N} H_p(X_i) - H_p(X_1, \ldots, X_N).$$

Here, the $H_p(X_i)$'s denote the marginal entropies and $H_p(X_1, \ldots, X_N)$ is the global entropy. This measure of stochastic interdependence of the units, which is called *multi-information*, is a generalization of the mutual information (see example in the introduction).

This article deals with the problem of finding natural low-dimensional exponential families that contain the maximizers of the multi-information in their closure. To this end we first consider a result on maximizers of the distance from an *arbitrary* exponential family [6], in the improved form obtained in [9]:

Proposition 3. Let $\mathcal{E}$ be an exponential family in $\mathcal{P}(\Omega)$ with dimension d. Then there exists an exponential family $\mathcal{E}^*$, $\mathcal{E} \subset \mathcal{E}^*$, with dimension less than or equal to $3d + 2$ such that the topological closure of $\mathcal{E}^*$ contains all local maximizers of $D_{\mathcal{E}}$.

This theorem is quite general, and is based on the observation that maximizers of the information divergence $D_{\mathcal{E}}$ have a reduced cardinality of their support, which is controlled by the dimension d of $\mathcal{E}$. The direct application of Proposition 3 of [9] to the exponential family $\mathcal{F}$ leads to the following statements on the local maximizers of the multi-information $I(X_1, \ldots, X_N) = D_{\mathcal{F}}$:

Corollary 3.1. There exists an exponential family $\mathcal{F}^*$ with

$$\dim \mathcal{F}^* \leq 3 \sum_{i=1}^{N} (n_i - 1) + 2 \leq 3N(n_N - 1) + 2$$

that contains all local maximizers of $I(X_1, \ldots, X_N)$ in its topological closure. In particular, in the binary case $n_i = 2$ for all i, $\dim \mathcal{F}^* \leq 3N + 2$.

In all such statements about exponential families over product spaces one should keep in mind, that the dimension of the exponential family $\mathcal{P}(\Omega_V)$ itself is of exponential growth in the number $N = |V|$ of units. So any exponential subfamily which is of polynomial growth in N is of large codimension.

Our main goal is now the following. Knowing about the existence of such low-dimensional exponential families $\mathcal{F}^*$, we want to analyze the relation between them and exponential families given by interaction structures between the N units.

More precisely, this article deals with the problem whether one can find low-dimensional exponential families $\mathcal{F}^*$ like in the Corollary 3.1 that are at the same time given by a low order of interaction. Before going into the details, we state an informal version of the main result of the paper (using terminology from statistical physics):

Informal Version of Theorem 5.1. If the cardinalities $n_1, \dots, n_N$ fulfill an inequality (see Theorem 4.4), the exponential family of pure pair-interactions (that is, pair-interactions without any external field) is sufficient for generating all global maximizers of the multi-information.

Let us have a closer look on this result for the binary case. In this case, we even find an exponential family of pure pair-interaction that has dimension $N - 1$, which is stronger than Corollary 3.1. More important, the pair-interactions form an explicit low dimensional exponential family that appears in many models in physics and biology (the units being called particles respectively neurons, the interactions fields resp. dendrites).

In Section 5, we will provide a rigorous formulation of our main result and prove it. This will be based on results concerning the structure of global maximizers of multi-information, which is discussed in the following Section 4.

4. THE STRUCTURE OF GLOBAL MAXIMIZERS OF MULTI–INFORMATION

4.1. General structure

Obviously, the maximal value of $I(X_1, \dots, X_N)$ is bounded as

$$I_p(X_1, \dots, X_N) = \sum_{i=1}^{N} H_p(X_i) - H_p(X_1, \dots, X_N) \le \sum_{i=1}^{N} \ln(n_i).$$

In fact, it turns out that in contrast to the quantum setting (see Remark 4.2 below), this upper bound is never reached. The following lemma gives an upper bound that is sharp in many interesting as well as important cases.

Lemma 4.1. Let p be a probability distribution on $\Omega_V = \Omega_1 \times \cdots \times \Omega_N$. Then:

$$I_p(X_1, \dots, X_N) \le \sum_{i=1}^{N-1} \ln(n_i). \tag{4.1}$$

Remark 4.2. With an orthonormal basis $f_1, \dots, f_n$ of the Hilbert space $\mathbb{C}^n$ we consider the (entangled) unit vector

$$\psi := \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \bigotimes_{i=1}^{N} f_k \in \bigotimes_{i=1}^{N} \mathbb{C}^n,$$

and the density operator ρ defined by the orthogonal projection onto the subspace spanned by ψ. In this setting, the mutual information is extended as

$$I(\rho) = \sum_{i=1}^{N} S(\rho_i) - S(\rho) = \operatorname{tr}\big(\rho \ln(\rho)\big) - \sum_{i=1}^{N} \operatorname{tr}\big(\rho_i \ln(\rho_i)\big)$$

where S denotes von Neumann entropy, and the ρ_i are the partial traces of ρ. As we see, this multi-information has the value $N \ln(n)$, which, according to Lemma 4.1, is not possible within the classical setting.

In the following, we consider the set

$$\mathcal{M}(\Omega_1, \dots, \Omega_N) := \left\{ p \in \overline{\mathcal{P}}(\Omega_V) \ : \ I_p(X_1, \dots, X_N) = \sum_{i=1}^{N-1} \ln(n_i) \right\}$$

of probability distributions that maximize, according to Lemma 4.1, in the case $\mathcal{M}(\Omega_1, \dots, \Omega_N) \neq \emptyset$ the multi-information $I(X_1, \dots, X_N)$. Up to isomorphism, everything depends only on the cardinalities $n_i = |\Omega_i|$ so that we sometimes write $\mathcal{M}(n_1, \dots, n_N)$ instead of $\mathcal{M}(\Omega_1, \dots, \Omega_N)$.

The next theorem characterizes the probability distributions in $\mathcal{M}(\Omega_1, \dots, \Omega_N)$.

Theorem 4.3. Let p be a probability distribution on Ω_V. Then $p \in \mathcal{M}(\Omega_1, \dots, \Omega_N)$ if and only if there exist a probability distribution $p^{(N)} \in \overline{\mathcal{P}}(\Omega_N)$ and surjective maps $\pi_i : \Omega_N \to \Omega_i$, $i = 1, \dots, N-1$, with

$$p^{(N)} \{\pi_i = \omega_i\} = \frac{1}{n_i} \qquad (\omega_i \in \Omega_i), \tag{4.2}$$

such that for all $(\omega_1, \dots, \omega_N) \in \Omega_V$

$$p(\omega_1, \dots, \omega_N) = \begin{cases} p^{(N)}(\omega_N), & \text{if } \omega_i = \pi_i(\omega_N),\ i = 1, \dots, N-1, \\ 0, & \text{otherwise.} \end{cases} \tag{4.3}$$

Theorem 4.3 allows us to say precisely under which conditions on the unit sizes n_i the theoretical maximum (4.1) of multi-information can be achieved (we use the shorthands $W := 2^{[N-1]} \backslash \{\emptyset\}$ and $n_A := (n_i)_{i \in A}$ and denote the greatest common divisor by GCD):

Theorem 4.4. We have $\mathcal{M}(\Omega_1, \ldots, \Omega_N) \neq \emptyset$ if and only if $n_N \geq n_{\min}$ for

$$n_{\min} = n_{\min}(n_1, \ldots, n_{N-1}) := \sum_{A \in W} (-1)^{|A|-1} \mathrm{GCD}(n_A).$$

Remarks 4.5.

1. In particular, $\mathcal{M}(\Omega_1, \ldots, \Omega_N) \neq \emptyset$ if
   (a) there are only $N = 2$ units, or
   (b) all units are identical ($n_1 = \ldots = n_N$).

   In the following Sections 4.2 and 4.3 we discuss these two important examples of Theorem 4.4 more precisely.

2. (a) We have the following inequalities for $n_{\min}$:

$$\max(n_1, \ldots, n_{N-1}) \leq n_{\min} \leq 1 + \sum_{i=1}^{N-1} (n_i - 1).$$

   These follow immediately from the defining relation $n_{\min} = |\bigcup_{i \in [N-1]} T_{n_i}|$ for $T_m := \{\frac{i}{m} \ : \ i \in [m]\}$, since $|T_{n_i}| = n_i$ and $1 \in T_{n_i}$.
   The left inequality becomes an equality iff the least common multiple $\mathrm{LCM}(n_{[N-1]}) = n_{N-1}$ (still assuming that $n_{i+1} \geq n_i$), whereas the right inequality becomes an equality iff the integers $n_1, \ldots, n_{N-1}$ are mutually prime.

   (b) Additionally, one gets

$$n_{\min} \leq \mathrm{LCM}(n_{[N-1]}) =: l,$$

   since for all $i \in [N-1]$ the inclusion $T_{n_i} \subset T_l$ holds true. Again we have equality iff $\mathrm{LCM}(n_{[N-1]}) = n_{N-1}$.

   (c) The global maximizers $p \in \mathcal{M}(\Omega_1, \ldots, \Omega_N)$ of multi-information that we construct simultaneously maximize the mutual information of the pairs $\{i, N\}$ of units.
   In the case $\mathrm{LCM}(n_{[N-1]}) = n_N$ they even simultaneously maximize the mutual information of all pairs $\{i, j\} \subset [N]$ of units.
   Both statements follow from direct inspection of p defined in (6.4).

4.2. The case of two units

We now discuss the case of two units, i. e. $N = 2$. In this case, the set

$$\mathcal{M}(\Omega_1, \Omega_2) = \{p \in \overline{\mathcal{P}}(\Omega_1 \times \Omega_2) : I_p(X_1, X_2) = \ln(n_1)\}$$

is non-empty and therefore consists of all global maximizers of the mutual information of the two units. We want to describe the structure of $\mathcal{M}(\Omega_1, \Omega_2)$ by stratifying it into a disjoint union of relatively open sets. In order to do that, we consider for $\Omega_1^* := \Omega_1 \cup \{0\}$ the following set of maps

$$\mathcal{S} := \{\pi : \Omega_2 \to \Omega_1^* : \pi(\Omega_2) \supset \Omega_1\}. \tag{4.4}$$

The relation

$$\sigma \preceq \pi \ :\Longleftrightarrow\ \sigma^{-1}(\omega_1) \subset \pi^{-1}(\omega_1) \text{ for all } \omega_1 \in \Omega_1$$

on $\mathcal{S}$ is a partial order which makes $\mathcal{S}$ a poset.

Example 4.6. For $\Omega_1 = \{1, 2\}$ and $\Omega_2 = \{1, 2, 3\}$ we get a poset $\mathcal{S}$ of 12 maps. The right graphics in Figure 2 shows the cover graph of the poset with vertex set $\mathcal{S}$. On the left we show the graphs of four of these maps. We have $\sigma \preceq \pi$ if σ is in the lower line and connected to π in the upper line (so-called Hasse diagram).

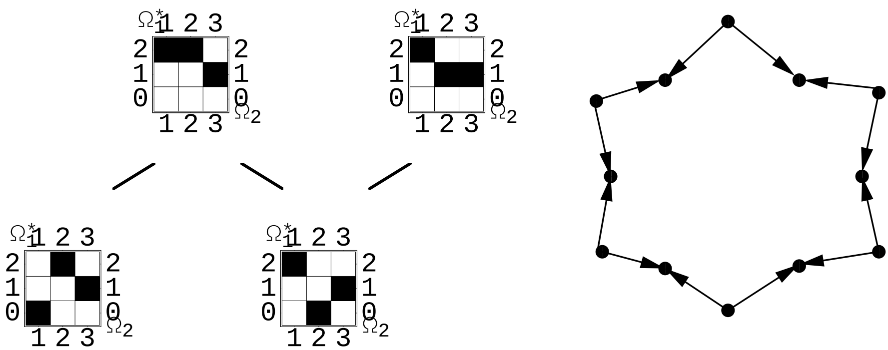

Fig. 2. The posets for $\Omega_1 = \{1, 2\}$, $\Omega_2 = \{1, 2, 3\}$.

We call a poset *connected* iff its cover graph is connected.

Lemma 4.7. The poset (4.4) is connected if and only if $n_1 < n_2$.

Given $\pi \in \mathcal{S}$ we consider the convex and relatively open set

$$\mathcal{M}_\pi(\Omega_1, \Omega_2) := \Big\{ p \in \overline{\mathcal{P}}(\Omega_1 \times \Omega_2) : \text{for all } \omega_1 \in \Omega_1, \sum_{\omega_2 \in \pi^{-1}(\omega_1)} p(\omega_1, \omega_2) = \frac{1}{n_1} \text{ and } p(\omega_1, \omega_2) > 0 \text{ iff } \pi(\omega_2) = \omega_1 \Big\}.$$

We denote by $S_{m,n}$ the Stirling numbers of the second kind (see for example [3]).

Theorem 4.8.

1. The set of global maximizers of the mutual information is a disjoint union

$$\mathcal{M}(\Omega_1, \Omega_2) = \biguplus_{\pi \in \mathcal{S}} \mathcal{M}_\pi(\Omega_1, \Omega_2)$$

   of sets $\mathcal{M}_\pi(\Omega_1, \Omega_2)$.

2. These sets have dimension

$$\dim \mathcal{M}_\pi(\Omega_1, \Omega_2) = |\pi^{-1}(\Omega_1)| - |\Omega_1|,$$

   and there are $n_1!\binom{n_2}{l} S_{l,n_1}$ sets $\mathcal{M}_\pi(\Omega_1, \Omega_2)$ of dimension $l - n_1$.

3. The inclusion $\mathcal{M}_\sigma(\Omega_1, \Omega_2) \subset \overline{\mathcal{M}_\pi}(\Omega_1, \Omega_2)$ holds if and only if $\sigma \preceq \pi$, and the set $\mathcal{M}(\Omega_1, \Omega_2)$ is connected if and only if $n_1 < n_2$.

Example 4.9. Continuing Example 4.6, for $n_1 = 2$ and $n_2 = 3$ the set $\mathcal{M}(2, 3)$ is the disjoint union of six points and six open intervals (see Figure 3, left), combined in the form of a hexagon (see Figure 3, right). So $\mathcal{M}(2, 3)$ is homeomorphic to S^1 in this case.

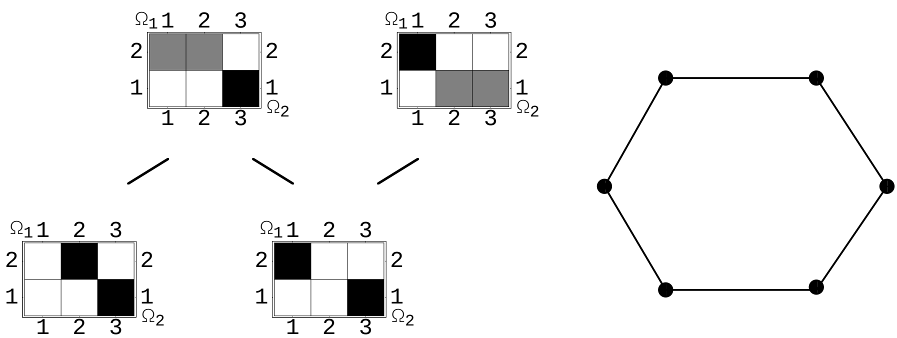

Fig. 3. The structure of $\mathcal{M}(2, 3)$.

4.3. The case of *N* equal units

This section deals with the important example of N units with $n_1 = \cdots = n_N =: n$. In that situation, Theorem 4.3 has the following direct implication.

Corollary 4.10. The set $\mathcal{M}(\Omega_1, \ldots, \Omega_N)$ consists of all probability distributions

$$\frac{1}{n} \sum_{\omega_N \in \Omega_N} \delta_{(\pi_1(\omega_N), \ldots, \pi_{N-1}(\omega_N), \omega_N)},$$

where $\pi_i : \Omega_N \to \Omega_i$, $i = 1, \ldots, N-1$, are one-to-one mappings. This implies

$$|\mathcal{M}(\Omega_1, \ldots, \Omega_N)| = (n!)^{N-1}, \tag{4.5}$$

and for all $p \in \mathcal{M}(\Omega_1, \ldots, \Omega_N)$,

$$I_p(X_1, \ldots, X_N) = (N-1) \cdot \ln(n),$$

$$|\operatorname{supp} p| = n. \tag{4.6}$$

Thus according to (4.5), the number of the maximizers of the multi-information grows exponentially in N. In particular, for binary units the set $\mathcal{M}(\Omega_1, \ldots, \Omega_N)$ has 2^{N-1} elements. In view of this fact, it is interesting that according to Corollary 3.1 there is an exponential family of dimension $\leq 3N + 2$ that approximates all these global maximizers of the multi-information. This bound can even be improved. Although it is not our main goal to do that we close this subsection by an interesting N-independent upper bound, which implies that for N binary units there exists an exponential family with dimension less than or equal to 5 that approximates all 2^{N-1} elements of $\mathcal{M}(\Omega_1, \ldots, \Omega_N)$.

Theorem 4.11. There exists an exponential family with dimension less than or equal to $(n^2 + 3n)/2$ that contains $\mathcal{M}(\Omega_1, \ldots, \Omega_N)$ in its closure.

This exponential family, however, is based on multibody interactions (in terms of statistical mechanics) between the units $i \in [N]$.

5. SUFFICIENCY OF LOW–ORDER INTERACTION FOR THE MAXIMIZATION OF MULTI–INFORMATION

Given a subset $A \subseteq [N] = \{1, \ldots, N\}$, we decompose $\omega \in \Omega_V$ in the form $\omega = (\omega_A, \omega_{[N] \setminus A})$ with $\omega_A \in \Omega_A$, $\omega_{[N]\setminus A} \in \Omega_{[N]\setminus A}$. We define $\mathcal{I}_A$ to be the subspace of functions that do not depend on the configurations $\omega_{[N]\setminus A}$:

$$\mathcal{I}_A := \Big\{ f \in \mathbb{R}^{\Omega_V} : \\ f(\omega_A, \omega_{V\setminus A}) = f(\omega_A, \omega'_{[N]\setminus A}) \text{ for all } \omega_A \in \Omega_A, \text{ and all } \omega_{[N]\setminus A}, \omega'_{[N]\setminus A} \in \Omega_{[N]\setminus A} \Big\}.$$

The orthogonal projection Π_A onto this $|\Omega_A|$-dimensional space with respect to the canonical scalar product

$$\langle f, g \rangle := \sum_{\omega \in \Omega_V} f(\omega) g(\omega) \qquad (f, g \in \mathbb{R}^{\Omega_V})$$

in $\mathbb{R}^{\Omega_V}$ is given by

$$\Pi_A(f)(\omega_A, \omega_{[N]\setminus A}) := \frac{1}{|\Omega_{[N]\setminus A}|} \sum_{\omega'_{[N]\setminus A} \in \Omega_{[N]\setminus A}} f(\omega_A, \omega'_{[N]\setminus A}).$$

In order to describe only the pure contributions of A to a function f, we "subtract" the contributions from subsets $B \subsetneq A$. This leads to the $\prod_{i\in A} \left(|\Omega_i| - 1\right)$-dimensional subspace

$$\widetilde{\mathcal{I}}_A := \mathcal{I}_A \cap \left(\bigcap_{B \subsetneq A} \mathcal{I}_B{}^{\perp} \right)$$

and the orthogonal decomposition $\mathbb{R}^{\Omega_V} = \bigoplus_{A\subseteq[N]} \widetilde{\mathcal{I}}_A$. Denoting the orthogonal projections onto $\widetilde{\mathcal{I}}_A$ by $\widetilde{\Pi}_A$ we thus have $\widetilde{\Pi}_A \widetilde{\Pi}_B = \delta_{A,B} \widetilde{\Pi}_A$ and

$$\Pi_A = \sum_{B\subseteq A} \widetilde{\Pi}_B, \qquad A \subseteq [N], \tag{5.1}$$

and every vector f has a unique representation as a sum of orthogonal vectors:

$$f = \sum_{A\subseteq[N]} \widetilde{\Pi}_A(f).$$

The f_A is called (*pure*) *interaction* among the units in A. With the Möbius inversion (5.1) implies

$$\begin{aligned} \widetilde{\Pi}_A(f) &= \sum_{B\subseteq A} (-1)^{|A\setminus B|} \Pi_B(f) \\ &= \sum_{B\subseteq A} (-1)^{|A\setminus B|} \frac{1}{|\Omega_{[N]\setminus B}|} \sum_{\omega'_{[N]\setminus B} \in \Omega_{[N]\setminus B}} f(\omega_B, \omega'_{[N]\setminus B}). \end{aligned}$$

Now we construct exponential families associated with such interaction spaces. The most general construction is based on a set of subsets of $[N]$. Given such a set $\boldsymbol{A} \subseteq 2^{[N]}$, we define the corresponding interaction space by

$$\widetilde{\mathcal{I}}_{\boldsymbol{A}} := \bigoplus_{A\in\boldsymbol{A}} \widetilde{\mathcal{I}}_A, \tag{5.2}$$

which generates the exponential family $\exp(\widetilde{\mathcal{I}}_{\boldsymbol{A}})$. We want to apply this definition to the more specific situation of interactions with fixed *order* k. Therefore, we define

$$\mathcal{I}^{(k)} := \widetilde{\mathcal{I}}_{\{A\subseteq[N]:|A|\le k\}}, \qquad \text{and} \qquad \widetilde{\mathcal{I}}^{(k)} := \widetilde{\mathcal{I}}_{\{A\subseteq[N]:|A|=k\}}.$$

We get the flag of vector spaces

$$\mathbb{R} \cong \mathcal{I}^{(0)} \subsetneq \mathcal{I}^{(1)} \subsetneq \mathcal{I}^{(2)} \cdots \subsetneq \mathcal{I}^{(N)} = \mathbb{R}^{\Omega_V},$$

and the corresponding hierarchy of exponential families

$$\exp(\mathcal{I}^{(0)}) \subsetneq \exp(\mathcal{I}^{(1)}) \subsetneq \exp(\mathcal{I}^{(2)}) \cdots \subsetneq \exp(\mathcal{I}^{(N)}) = \mathcal{P}(\Omega_V),$$

Here, $\exp(\mathcal{I}^{(0)})$ contains exactly one element, namely the center of the simplex.

The exponential family $\exp(\mathcal{I}^{(1)})$ is nothing but the exponential family $\mathcal{F}$ of factorizable distributions. Thus, the multi-information vanishes exactly on the topological closure of $\exp(\mathcal{I}^{(1)})$.

Now we determine for a nonempty set $\mathcal{M}(\Omega_1, \ldots, \Omega_N)$ of maximizers the lowest order k such that $\mathcal{M}(\Omega_1, \ldots, \Omega_N)$ is contained in the topological closure of $\exp(\mathcal{I}^{(k)})$. The first possible candidate for this is given by $k = 2$. The following theorem states that this is also sufficient.

Theorem 5.1. There exists an exponential family $\mathcal{F}^* \subseteq \exp(\widetilde{\mathcal{I}}^{(2)})$ of dimension $\dim(\mathcal{F}^*) = (n_N - 1)\sum_{i=1}^{N-1}(n_i - 1)$ containing in its closure all global maximizers of the multi-information ($\mathcal{M}(\Omega_1, \ldots, \Omega_N) \subset \overline{\mathcal{F}^*}$).

This theorem represents our main result which we already stated informally in Section 3. Note that compared with Theorem 4.11 for large N Theorem 5.1 leads to an exponential family $\mathcal{F}^*$ of higher dimension. On the other hand, we still have an exponential (in N) codimension in the simplex $\overline{\mathcal{P}}(\Omega_V)$.

In addition to that, the exponential family of Theorem 5.1 represents a concrete model that appears in many applications in physics and biology. For instance, within the field of neural networks, the exponential family $\exp(\mathcal{I}^{(2)})$, which contains $\exp(\widetilde{\mathcal{I}}^{(2)})$ as a subfamily, is known as the family of Boltzmann machines, [1, 2, 5]. Applied to this context, our result states that Boltzmann machines are able to generate all distributions that have globally maximal multi-information, and that their dimensionality $\binom{N}{2}$ is not minimal for $N > 2$.

Examples 5.2.

1. **The Case of Two Units.** In this case, the hierarchy of interactions ends with $k = 2$, because we have just two units. Thus the simplex $\mathcal{P}(\Omega_1 \times \Omega_2)$ is equal to the exponential family $\exp(\mathcal{I}^{(2)})$, which has dimension $n_1 n_2 - 1$. The codimension of the subfamily $\exp(\widetilde{\mathcal{I}}^{(2)})$ of Theorem 5.1 then is $n_1 + n_2 - 2$. Applied to our example of two binary units from the introduction, we see that

$$\dim(\exp(\widetilde{\mathcal{I}}^{(2)})) = 1$$

In Figure 1, we obtain this family by simply taking the convex combinations of the two maximizers:

$$\exp(\widetilde{\mathcal{I}}^{(2)}) = \left\{ \frac{1-\lambda}{2}\left(\delta_{(0,0)} + \delta_{(1,1)}\right) + \frac{\lambda}{2}\left(\delta_{(1,0)} + \delta_{(0,1)}\right) : 0 < \lambda < 1 \right\}.$$

2. **The Case of N Equal Units.** According to Theorem 4.3 for $|\Omega_i| = n$ we have $|\mathcal{M}(\Omega_1, \dots, \Omega_N)| = (n!)^{N-1}$ maximizers, which are, according to Theorem 5.1, contained in the closure of an exponential family $\mathcal{F}^*$ of pure pair interactions, with
$$\dim(\mathcal{F}^*) = (N-1)(n-1)^2.$$

6. PROOFS

We fix the following notations: For $V' \subset [N]$, $H_{V'}$ denotes the entropy of the random variable $X_{V'}$. Obviously $H_V = H$, and $H_{\{i\}} = H_i$. For two subsets $V', V'' \subset [N]$, $H_{(V'' \mid V')}$ is the conditional entropy of $X_{V''}$ given $X_{V'}$. For $V' = \{a_1, \dots, a_L\}$ and $V'' = \{b_1, \dots, b_M\}$ we also write $H_{(b_1,\dots,b_M \mid a_1,\dots,a_L)}$ instead of $H_{(V'' \mid V')} = H_{(\{b_1,\dots,b_M\} \mid \{a_1,\dots,a_L\})}$. Now let $V_1, \dots, V_r$ be a set of disjoint subsets of $[N] = \{1, \dots, N\}$. The multi-information of these subsystems is given by $I_{\{V_1,\dots,V_r\}} = \sum_{j=1}^r H_{V_j} - H_{V_1 \uplus \dots \uplus V_r}$. In the case where the subsets of $[N]$ have cardinality one, we also write $I_{\{i_1,\dots,i_r\}}$ instead of $I_{\{\{i_1\},\dots,\{i_r\}\}}$. We obviously have $I_V = I$.

P r o o f o f L e m m a 4.1. By the chain rule $H(X,Y) = H(X) + H(Y \mid X)$

$$\begin{aligned} I_p(X_1, \dots, X_N) &= \sum_{i=1}^N H_p(X_i) - H_p(X_1, \dots, X_N) \\ &= \sum_{i=1}^{N-1} H_p(X_i) - \big(H_p(X_1, \dots, X_N) - H_p(X_N)\big) \\ &= \sum_{i=1}^{N-1} H_p(X_i) - H_p(X_1, \dots, X_{N-1} \mid X_N) \le \sum_{i=1}^{N-1} H_p(X_i) \le \sum_{i=1}^{N-1} \ln(n_i), \end{aligned}$$

proving the lemma. □

P r o o f o f T h e o r e m 4.3. If a probability distribution p on Ω_V has the form (4.3) with a distribution $p^{(N)} \in \overline{\mathcal{P}}(\Omega_N)$ and surjective maps $\pi_i : \Omega_N \to \Omega_i$ that satisfy (4.2), then $I(p) = \sum_{i=1}^{N-1} \ln(n_i)$:

$$\begin{aligned} I(p) &= \sum_{i=1}^N H_i(p) - H(p) \\ &= \sum_{i=1}^N H_i(p) - H_N(p) \underbrace{- H_{(1|N)}(p) - H_{(2|1,N)}(p) - \dots - H_{(N-1|1,2,\dots,N-2,N)}(p)}_{=0} \\ &= \sum_{i=1}^{N-1} \ln(n_i). \end{aligned}$$

Now we prove the opposite implication. Therefore we assume $I(p) = \sum_{i=1}^{N-1} \ln(n_i)$. This gives us

$$H_i(p) = \ln(n_i) \qquad (i = 1, \dots, N-1). \tag{6.1}$$

Otherwise the existence of an $i_0 \in \{1, \dots, N-1\}$ with $H_{i_0}(p) < \ln(n_{i_0})$ would imply the following contradiction

$$\begin{aligned} I(p) &= \sum_{i=1}^{N} H_i(p) - H(p) \\ &= \sum_{i=1}^{N-1} H_i(p) + H_N(p) - \big(H_N(p) + H_{(1,\dots,N-1\,|\,N)}(p)\big) \\ &\le \sum_{\substack{i=1 \\ i \neq i_0}}^{N-1} H_i(p) + H_{i_0}(p) < \sum_{i=1}^{N-1} \ln(n_i). \end{aligned}$$

From (6.1) we have

$$H(p) = \sum_{i=1}^{N} H_i(p) - I(p) = \left(\sum_{i=1}^{N-1} \ln(n_i) + H_N(p) \right) - \sum_{i=1}^{N-1} \ln(n_i) = H_N(p). \tag{6.2}$$

Now we set $p^{(N)} := p_N$, and define a Markov kernel $K : (\Omega_1 \times \cdots \times \Omega_{N-1}) \times \Omega_N \to [0, 1]$ by

$$K(\omega_1, \dots, \omega_{N-1} \,|\, \omega_N) := \begin{cases} \frac{p(\omega_1,\dots,\omega_N)}{p_N(\omega_N)}, & \text{if } p_N(\omega_N) > 0 \\ \frac{1}{n_1 \cdots n_{N-1}}, & \text{if } p_N(\omega_N) = 0\text{š}. \end{cases}$$

In these definitions we get

$$\begin{aligned} &H(p) - H_N(p) \\ &= \sum_{\substack{\omega_N \in \Omega_N \\ p_N(\omega_N) > 0}} p_N(\omega_N) \Bigg(\ln p_N(\omega_N) \\ &\qquad - \sum_{\substack{(\omega_1,\dots,\omega_{N-1}) \in \\ \Omega_1 \times \cdots \times \Omega_{N-1}}} K(\omega_1, \dots, \omega_{N-1} \,|\, \omega_N) \ln \Big(p_N(\omega_N)\, K(\omega_1, \dots, \omega_{N-1} \,|\, \omega_N) \Big) \Bigg) \\ &= \sum_{\substack{\omega_N \in \Omega_N \\ p_N(\omega_N) > 0}} p_N(\omega_N)\, H\big(K(\cdot \,|\, \omega_N)\big) \ge 0. \end{aligned}$$

From (6.2) this implies $H\big(K(\cdot \,|\, \omega_N)\big) = 0$ for all ω_N with $p_N(\omega_N) > 0$. This implies the existence of maps $\pi_i : \Omega_N \to \Omega_i$ with

$$p(\omega_1, \dots, \omega_N) = p^{(N)}(\omega_N) \prod_{i=1}^{N-1} \delta_{\omega_i, \pi_i(\omega_N)}.$$

Because of $H_i(p) = \ln(n_i)$ for all $i \in \{1, \dots, N-1\}$, these maps must be surjective. □

Proof of Theorem 4.4.

Proof that $\mathcal{M}(\Omega_1, \dots, \Omega_N) \neq \emptyset$ if $n_N \geq n_{\min}$:
For $m \in \mathbb{N}$ set $T_m := \left\{ \frac{i}{m} \ : \ i \in [m] \right\}$. We claim that the cardinality of

$$T_\Omega := \bigcup_{i \in [N-1]} T_{n_i}$$

is given by $|T_\Omega| = n_{\min}$. This follows by the inclusion-exclusion principle if

$$\left| \bigcap_{i \in A} T_{n_i} \right| = \mathrm{GCD}(n_A) \qquad (A \in W), \tag{6.3}$$

since

$$\left| \bigcup_{i \in [N-1]} T_{n_i} \right| = \sum_{A \in W} (-1)^{|A|-1} \left| \bigcap_{i \in A} T_{n_i} \right| .$$

To prove (6.3), we set $m_A := \mathrm{GCD}(n_A)$ and note that $T_{n_i} \supseteq T_{m_A} \quad (i \in A)$. Thus $\left| \bigcap_{i \in A} T_{n_i} \right| \geq |T_{m_A}| = m_A$.

To show the converse inequality $\left| \bigcap_{i \in A} T_{n_i} \right| \leq m_A$ we note that for some $\tilde{m} \in \mathbb{N}$ we have $\bigcap_{i \in A} T_{n_i} = T_{\tilde{m}}$. Thus for all $i \in A$ there exist $\ell_i \in [n_i]$ with $\frac{\ell_i}{n_i} = \frac{1}{\tilde{m}} = \min(T_{\tilde{m}})$, or $n_i = \ell_i \tilde{m}$. Thus $\tilde{m}$ divides all $n_i \quad (i \in A)$ and – being the largest such integer – equals $m_A = \mathrm{GCD}(n_A)$.

Now we write T_{Ω_V} in the form $\{d_1, \dots, d_{n_{\min}}\}$ and set $d_0 := 0$, with ordering $d_i > d_{i-1} \quad (i \in [n_{\min}])$. The map

$$\Phi : T_{\Omega_V} \to \Omega_V, \qquad \Phi(d_j)_i := \begin{cases} \lceil d_j n_i \rceil, & i \in [N-1] \\ j, & i = N \end{cases}$$

is well defined, since $\lceil d_j n_i \rceil \in [n_i] \quad (i \in [N-1])$, and by our assumption $n_N \geq n_{\min}$ which implies $j \in [n_N]$. The function

$$p : \Omega_V \to \mathbb{R}, \qquad p := \sum_{j=1}^{n_{\min}} (d_j - d_{j-1}) \delta_{\Phi(d_j)} \tag{6.4}$$

is a probability distribution since $d_j - d_{j-1} > 0$ and

$$\sum_{j=1}^{n_{\min}} (d_j - d_{j-1}) = d_{n_{\min}} - d_0 = 1.$$

For all $i \in [N-1]$ and $\ell \in [n_i]$ the ith marginal probability equals

$$\begin{aligned} p_i(\ell) &= \sum_{\omega \in \times_{j \in [N] \setminus \{i\}} n_j} p(\ell, \omega) = \sum_{\substack{\omega_N \in [n_{\min}] \\ \lceil d_{\omega_N} n_i \rceil = \ell}} (d_{\omega_N} - d_{\omega_N - 1}) \\ &= \sum_{j : d_j \in \left(\frac{\ell-1}{n_i}, \frac{\ell}{n_i} \right]} (d_j - d_{j-1}) = \frac{\ell}{n_i} - \frac{\ell - 1}{n_i} = \frac{1}{n_i}. \end{aligned}$$

We thus meet the condition of Theorem 4.3 showing that $p \in \mathcal{M}(\Omega_1, \dots, \Omega_N)$.

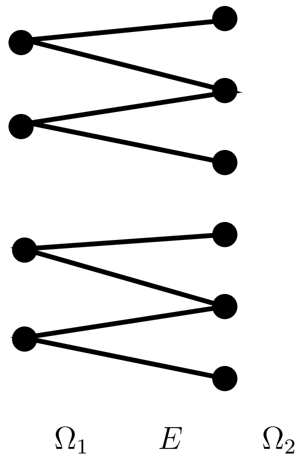

Fig. 4. A bipartite graph $(\tilde{V}, G)$ for $N+1 = 3$ units, with $|\Omega_1| = 4$, $|\Omega_1| = 6$, $|\Omega_3| \geq 8$ and a maximizer $p \in \mathcal{M}(\Omega_1, \Omega_2, \Omega_3)$. $(\tilde{V}, G)$ has the two components C_1, C_2.

Proof that $\mathcal{M}(\Omega_1, \ldots, \Omega_N) = \emptyset$ *if* $n_N < n_{\min}$*:*

- The statement is trivial for $N = 2$ (remember that we assume $n_{i+1} \geq n_i$). Assume now that it is proven for all product spaces of at most $N \in \mathbb{N}$ units. Then for a probability distribution $p \in \mathcal{M}(\Omega_1, \ldots, \Omega_{N+1})$ consider its marginal $\tilde{p} \in \overline{\mathcal{P}}(\Omega_{[N]})$.

  We associate to $\tilde{p}$ a N-partite graph $(\tilde{V}, E)$ whose vertex set is the disjoint union $\tilde{V} := \dot{\bigcup}_{i=1}^{N} \Omega_i$. To every $\omega = (\omega_1, \ldots, \omega_N) \in \mathrm{supp}(\tilde{p}) \subseteq \Omega_{[N]}$ belongs the complete graph on the vertex set $\{\omega_1, \ldots, \omega_N\} \subset \tilde{V}$ with edge set $G_\omega := \big\{\{\omega_i, \omega_j\} \subset \tilde{V} \mid 1 \leq i < j \leq N\big\}$ on the N vertices $\omega_1, \ldots, \omega_N$. Then the edge set
$$E := \bigcup_{\omega \in \mathrm{supp}(\tilde{p})} G_\omega$$
  on $\tilde{V}$ is indeed N-partite. By the strict positivity (4.2) of the p-marginals no vertex $v \in \tilde{V}$ is isolated.

- Every edge set $G_\omega \subseteq E$ is contained in the induced subgraph of exactly one connected component $C \subseteq \tilde{V}$ of the graph $(\tilde{V}, E)$. We attribute to G_ω the weight $\tilde{p}(\omega)$, and to a connected component C of the graph $(\tilde{V}, E)$ the sum of the weights of the G_ω contained in it.

  These weights $w(C)$ of the connected components C are not arbitrary numbers in $(0, 1]$. Instead, we know from Theorem 4.3 that the marginal distributions $p_i : \Omega_i \to [0, 1]$ of p (and thus of $\tilde{p}$, too) have the Laplace form
$$p_i(\omega_i) = \frac{1}{n_i} \qquad (i \in [N],\ \omega_i \in \Omega_i).$$

Therefore $w(C)$ is simultaneously an integer multiple of $1/n_i \quad (i \in [N])$ and thus an integer multiple of $\mathrm{GCD}(n_{[N]})$. This implies the upper bound $\mathrm{GCD}(n_{[N]})$ for the number of connected components C of the N-partite graph $(\tilde{V}, E)$.

- For the case of $N+1=3$ units this already suffices to show the bound $n_3 \geq n_{\min} = n_1 + n_2 - \mathrm{GCD}(n_1, n_2)$. In this case the complete graphs are of cardinality $|G_\omega| = (N-1)! = 1$ so that $|E| = |\mathrm{supp}(\tilde{p})|$.

  In general a graph on a vertex set of $v \in \mathbb{N}$ vertices with $e \in \mathbb{N}_0$ edges has at least $\max(v-e, 1)$ connected components. In the case at hand $v = n_1 + n_2$, and there are at most $\mathrm{GCD}(n_1, n_2)$ connected components. So

$$\begin{aligned} n_3 &\geq |\mathrm{supp}(p)| \geq |\mathrm{supp}(\tilde{p})| = |E| = e \\ &\geq v - c = (n_1 + n_2) - \mathrm{GCD}(n_1, n_2) = n_{\min}. \end{aligned}$$

- For arbitrary $N+1>3$ this argument must be modified, since then $|G_\omega| = (N-1)! > 1$.

  First of all we can substitute G_ω by any spannning tree $T_\omega \subset G_\omega$, and still the connected components C' of $(\tilde{V}, E')$ with $E' := \bigcup_{\omega \in \mathrm{supp}(\tilde{p})} T_\omega$ coincide with the connected components C of $(\tilde{V}, E)$. Each of these spanning trees has only $|T_\omega| = N-1$ edges. However in general E', too is not a disjoint union of the T_ω.

  We thus decompose the set $\mathrm{supp}(\tilde{p})$ into a disjoint union

$$\mathrm{supp}(\tilde{p}) = \bigcup_{k=1}^{N} A_k, \tag{6.5}$$

  beginning with an arbitrarily chosen set A_N of representatives $\omega \in C$ of the connected components $C \subseteq \Omega_{[N]}$. The estimate on the number of these components implies $|A_N| \geq \mathrm{GCD}(n_{[N]})$, and for $\omega \neq \omega' \in A_N$ the edge sets G_ω and $G_{\omega'}$ are disjoint.

  Next we arrange the elements $\omega' \in C$ of the connected component C containing $\omega \in A_N$ in the form of a spanning tree, with $G_{\omega'} \cap G_{\omega''} \neq \emptyset$ for $\{\omega', \omega''\}$ being an edge of that tree. For $\omega' = (\omega'_1, \ldots, \omega'_N) \in C$ of distance $d(\omega')$ from $\omega \in A_N$ we put $\omega' \in A_k$ if there are exactly k indices $i \in [N]$ with ω'_i not being equal to any ω''_i for $\omega'' = (\omega''_1, \ldots, \omega''_N)$ with $d(\omega'') < d(\omega')$. This indeed gives a partition of the form (6.5).

  Then by our induction hypothesis

$$|A_k| \geq \sum_{\substack{B \subseteq [N] \\ |B| \geq k}} (-1)^{|B|-k} \binom{|B|}{k} \mathrm{GCD}(n_B) \qquad (k = 1, \ldots, N). \tag{6.6}$$

  Namely for $k = N$ (6.6) reduces to $|A_N| \geq \mathrm{GCD}(n_{[N]})$ which has been shown to be true. So if (6.6) would not hold, for the smallest $k < N$ violating

(6.6), we would find a $B \subseteq [N]$ of cardinality $|B| = k < N$, whose marginal distribution p_B has support of cardinality $\hat{n}_{k+1} := |\mathrm{supp}(p_B)| < n_{\min}(B) = \sum_{\substack{\tilde{B} \subseteq B \\ \tilde{B} \neq \emptyset}} (-1)^{|\tilde{B}|-1} \mathrm{GCD}(n_{\tilde{B}})$, see (6.7) below.

But this would contradict our induction assumption, since then the system $\hat{\Omega} := \left(\times_{i \in B} [n_i] \right) \times [\hat{n}_{k+1}]$ would have the optimizing probability distribution

$$\hat{p} : \hat{\Omega} \to [0,1], \qquad \hat{p}(\omega_B, l) := \delta_{e(l),\omega_B} \, p_B(\omega_B)$$

for some bijection $e : [\hat{n}_{k+1}] \to \mathrm{supp}(p_B)$, but yet not meet the criterium $\hat{n}_{k+1} \geq n_{\min}(B)$.

Summing the cardinalities (6.6), we obtain

$$\begin{aligned} |\mathrm{supp}(\tilde{p})| &= \sum_{k=1}^{N} |A_k| \geq \sum_{k=1}^{[N]} \sum_{\substack{B \subseteq [N] \\ |B| \geq k}} (-1)^{|B|-k} \binom{|B|}{k} \mathrm{GCD}(n_B) \\ &= \sum_{\substack{B \subseteq [N] \\ B \neq \emptyset}} (-1)^{|B|} \mathrm{GCD}(n_B) \sum_{k=1}^{|B|} (-1)^k \binom{|B|}{k} \\ &= \sum_{\substack{B \subseteq [N] \\ B \neq \emptyset}} (-1)^{|B|-1} \mathrm{GCD}(n_B) = n_{\min}, \end{aligned} \tag{6.7}$$

which is the induction step. □

P r o o f o f L e m m a 4.7. If $n_1 = n_2$ then the maps $\pi \in \mathcal{S}$ are isomorphisms $\pi : \Omega_2 \to \Omega_1$, so that $\sigma \preceq \pi$ only for $\sigma = \pi$. Thus in that case $\mathcal{S}$ is connected iff $|\mathcal{S}| = 1$, i. e. $n_1 = n_2 = 1$. This contradicts our assumption $n_1, n_2 \geq 2$.

If $n_2 > n_1$ and $|\pi^{-1}(\omega_1)| > 1$ for $\pi \in \mathcal{S}$ and some $\omega_1 \in \Omega_1$, say $\pi(\omega_2') = \omega_1$, then $\sigma \preceq \pi$ for

$$\sigma \in \mathcal{S}, \qquad \sigma(\omega_2) := \begin{cases} \pi(\omega_2), & \text{if } \omega_2 \neq \omega_2' \\ 0, & \text{if } \omega_2 = \omega_2'. \end{cases}$$

So we need only show that any $\pi', \pi'' \in S$ which are injective onto Ω_1 are indeed connected.

1. In the first step we move π' along the poset graph in order to decrease the cardinality of the symmetric difference $(\pi')^{-1}(0) \Delta (\pi'')^{-1}(0)$. So we assume that there exist

$$\omega' \in (\pi')^{-1}(0) \backslash (\pi'')^{-1}(0) \quad \text{and} \quad \omega'' \in (\pi'')^{-1}(0) \backslash (\pi')^{-1}(0)$$

and set

$$\pi \in \mathcal{S}, \qquad \pi(\omega) := \begin{cases} 0, & \text{if } \omega = \omega'' \\ \pi'(\omega''), & \text{if } \omega = \omega' \\ \pi'(\omega), & \text{otherwise.} \end{cases}$$

Both π' and π are covered by

$$\rho \in \mathcal{S}, \qquad \rho(\omega) := \begin{cases} \pi'(\omega''), & \text{if } \omega = \omega' \\ \pi'(\omega), & \text{otherwise,} \end{cases}$$

and

$$|\pi^{-1}(0)\Delta(\pi'')^{-1}(0)| = |(\pi')^{-1}(0)\Delta(\pi'')^{-1}(0)| - 2.$$

By iterating the argument we can assume w.l.o.g. that $(\pi')^{-1}(0) = (\pi'')^{-1}(0)$.

2. In fact it is sufficient to treat the case where the permutation

$$\pi'' \circ (\pi')^{-1}\,|_{\Omega_1} : \Omega_1 \to \Omega_1$$

is a transposition, as the transpositions generate the symmetric group. So there exist $\omega^I \neq \omega^{II} \in \Omega_2$ with

$$\pi''(\omega) = \begin{cases} \pi'(\omega^I), & \text{if } \omega = \omega^{II} \\ \pi'(\omega^{II}), & \text{if } \omega = \omega^I \\ \pi'(\omega), & \text{otherwise,} \end{cases}$$

and we choose $\hat{\omega} \in \Omega_2$ so that $\pi'(\hat{\omega}) = \pi''(\hat{\omega}) = 0$.

Defining $\rho, \rho'' \in \mathcal{S}$ by

$$\rho'(\omega) := \begin{cases} \pi'(\omega^{II}), & \text{if } \omega = \hat{\omega} \\ 0, & \text{if } \omega = \omega^{II} \\ \pi'(\omega), & \text{otherwise} \end{cases} \quad \text{resp.} \quad \rho''(\omega) := \begin{cases} \pi''(\omega^I), & \text{if } \omega = \hat{\omega} \\ 0, & \text{if } \omega = \omega^I \\ \pi''(\omega), & \text{otherwise,} \end{cases}$$

π' and ρ' are covered by $\sigma' \in \mathcal{S}$ and similarly π'' and ρ'' are covered by $\sigma'' \in \mathcal{S}$ with

$$\sigma'(\omega) := \begin{cases} \pi'(\omega^{II}), & \text{if } \omega = \hat{\omega} \\ \pi'(\omega), & \text{otherwise} \end{cases} \quad \text{resp.} \quad \sigma''(\omega) := \begin{cases} \pi''(\omega^I), & \text{if } \omega = \hat{\omega} \\ \pi''(\omega), & \text{otherwise.} \end{cases}$$

Now as $\pi'(\omega^{II}) = \pi''(\omega^I)$, both ρ' and ρ'' are covered by

$$\tau \in \mathcal{S}, \qquad \tau(\omega) := \begin{cases} \pi'(\omega^{II}), & \text{if } \omega = \hat{\omega} \\ \pi'(\omega^I), & \text{if } \omega = \omega^{II} \\ \pi'(\omega) & \text{otherwise.} \end{cases}$$

This shows that the poset graph is connected. □

Proof of Theorem 4.8. To simplify notation, we set $\mathcal{M} := \mathcal{M}(\Omega_1, \Omega_2)$, and $\mathcal{M}_\pi := \mathcal{M}_\pi(\Omega_1, \Omega_2)$ for $\pi \in \mathcal{S}$.

1. We have $\mathcal{M}_\pi \subset \mathcal{M}$ since for the elements of $\mathcal{M}_\pi$ the characterisation of Theorem 4.3 holds true. Furthermore for $\sigma, \pi \in \mathcal{S}$ with $\sigma \neq \pi$ there exists $(\omega_2, \omega_1) \in \text{graph}(\pi)$ with $(\omega_2, \omega_1) \notin \text{graph}(\sigma)$ or vice versa. Thus for $p \in \mathcal{M}_\pi$ we have $p(\omega_1, \omega_2) > 0$ but for $p \in \mathcal{M}_\sigma$ we have $p(\omega_1, \omega_2) = 0$ showing that $\mathcal{M}_\pi \cap \mathcal{M}_\sigma = \emptyset$.

   Finally for $p \in \mathcal{M}$ by Theorem 4.3 there exists a surjective map $\tilde{\pi} : \Omega_2 \to \Omega_1$ with $p(\omega_1, \omega_2) = 0$ whenever $\tilde{\pi}(\omega_2) \neq \omega_1$. Given $\tilde{\pi}$, we construct $\pi \in \mathcal{S}$ by setting
$$\pi(\omega_2) := \begin{cases} \tilde{\pi}(\omega_2), & \text{if } p(\tilde{\pi}(\omega_2), \omega_2) > 0 \\ 0, & \text{if } p(\tilde{\pi}(\omega_2), \omega_2) = 0. \end{cases}$$
   As by Theorem 4.3 we have $\sum_{\omega_2 \in \tilde{\pi}^{-1}(\omega_1)} p(\omega_1, \omega_2) = \frac{1}{n_1} > 0$, the function $\pi : \Omega_2 \to \Omega_1^*$ so constructed has the property $\pi(\Omega_2) \supset \Omega_1$ making it an element of $\mathcal{S}$.

2. Given $\omega_1 \in \Omega_1$, the simplex of $|\pi^{-1}(\omega_1)|$ numbers $p(\omega_1, \omega_2) > 0$ with $\omega_2 \in \pi^{-1}(\omega_1)$ meeting $\sum_{\omega_2 \in \pi^{-1}(\omega_1)} p(\omega_1, \omega_2) = \frac{1}{n_1}$ has dimension $|\pi^{-1}(\omega_1)| - 1$, implying the formula for $\dim \mathcal{M}_\pi$.

   If $\dim \mathcal{M}_\pi = l - n_1$, the surjective map $\hat{\pi} : \hat{\Omega}_2 \to \Omega_1$ with $\hat{\Omega}_2 := \pi^{-1}(\Omega_1) \subset \Omega_2$ and $\hat{\pi} := \pi \mid_{\hat{\Omega}_2}$ is defined on a subset $\hat{\Omega}_2 \subset \Omega_2$ of size l. There are precisely $\binom{n_2}{l}$ such subsets, and there are precisely $n_1! S_{l,n_1}$ such surjective maps from $\hat{\Omega}_2$ onto Ω_1, see Aigner [3], Chapter 3.1.

   (3) If $n_1 = n_2$ then $\mathcal{S}$ coincides with the set of bijections $\pi : \Omega_2 \to \Omega_1$, and $|\mathcal{M}_\pi| = 1$. Thus in this case $\mathcal{M}$ is not connected for $n_1 \geq 2$. If, however $n_2 > n_1$, the poset $\mathcal{S}$, seen as a graph, is connected.

   The topological closure of $\mathcal{M}_\pi$ is given by
$$\overline{\mathcal{M}}_\pi = \left\{ p \in \overline{\mathcal{P}}(\Omega_1 \times \Omega_2) : \sum_{\omega_2 \in \pi^{-1}(\omega_1)} p(\omega_1, \omega_2) = \frac{1}{n_1}, p(\omega_1, \omega_2) = 0 \text{ if } \pi(\omega_2) \neq \omega_1 \right\}.$$
   Thus $\overline{\mathcal{M}}_\pi = \biguplus_{\sigma \preceq \pi} \mathcal{M}_\sigma$. □

Proof of Corollary 4.10. All statements directly follow from Theorem 4.3. □

Proof of Theorem 4.11. We choose a map $\phi = (\phi_1, \ldots, \phi_n) : \Omega_V \to \mathbb{R}^n$ such that the points $\phi(\omega)$, $\omega \in \Omega_V$, are in general position; that is, each k elements of $\phi(\Omega_V)$ with $k \leq n+1$ are affinely independent. This property guarantees that for each set $\Sigma \subset \Omega_V$, $|\Sigma| = n$, there exist real numbers $a_1, \ldots, a_n, b$ such that
$$\left\{ \omega \in \Omega_V : \sum_{i=1}^n a_i \, \phi_i(\omega) = b \right\} = \Sigma \tag{6.8}$$
holds. We consider the exponential family $\mathcal{G}^*$ that is generated by c and
$$\phi_1, \ldots, \phi_n, \qquad \phi_i \, \phi_j \quad (1 \leq i \leq j \leq n).$$

We have

$$\dim \mathcal{G}^* \leq \frac{n^2+3n}{2}.$$

Now let p be an element of $\mathcal{M}(N \times n)$. From Theorem 4.10 we know that $|\mathrm{supp}\, p| = n$. We prove that there exists a sequence in $\mathcal{G}^*$ that converges to p. We choose a sequence $\beta_m \uparrow \infty$ and real numbers $a_1, \dots, a_n, b$ satisfying (6.8) with $\Sigma = \mathrm{supp}\, p$. Then with

$$E^{(m)} := -\beta_m \left(\sum_{i=1}^{n} a_i \, \phi_i - b \right)^2,$$

the sequence

$$\frac{\exp E^{(m)}}{\sum_{\omega' \in \Omega_V} \exp E^{(m)}(\omega')} \in \mathcal{G}^*$$

converges to p. □

P r o o f o f T h e o r e m 5.1. Using def. (5.2), we consider for $\boldsymbol{A} := \{\{1,N\},\{2,N\}, \dots, \{N-1,N\}\} \subset 2^{[N]}$ the linear subspace

$$\widetilde{\mathcal{I}}_{\boldsymbol{A}} \subset \widehat{\mathcal{I}}_2$$

of pure pair interactions of the Nth unit with all other units. The exponential family $\mathcal{F}^* := \exp(\widetilde{\mathcal{I}}_{\boldsymbol{A}}) \subset \mathcal{P}(\Omega_V)$ is of dimension

$$\dim(\mathcal{F}^*) = (n_N - 1) \sum_{i=1}^{N-1} (n_i - 1),$$

as asserted in Theorem 5.1.

Given a maximizer $p \in \mathcal{M}(\Omega_1, \dots, \Omega_N)$, we now construct a sequence of probability distributions

$$q^{(m)} := \exp(\tilde{f}^{(m)}) \in \mathcal{F}^* \qquad (m \in \mathbb{N})$$

and show that $\lim_{m \to \infty} q^{(m)} = p$.

Here the functions $\tilde{f}^{(m)} \in \widetilde{\mathcal{I}}_{\boldsymbol{A}}$ are defined as the orthogonal projections onto $\widetilde{\mathcal{I}}_{\boldsymbol{A}}$ of $f^{(m)} \in \mathcal{I}^{(2)}$

$$f^{(m)}(\omega) := \prod_{i=1}^{N-1} \delta_{\omega_i, \pi_i(\omega_N)} \, \frac{m + \ln\left(p^{(N)}(\omega_N) + 1/m\right)}{N-1} \qquad (\omega \in \Omega_V).$$

For $\omega, \omega' \in \mathrm{supp}(p)$

$$\begin{aligned} \frac{q^{(m)}(\omega)}{q^{(m)}(\omega')} &= \exp\left(\left[m + \ln\left(p^{(N)}(\omega_N) + \frac{1}{m}\right)\right] - \left[m + \ln\left(p^{(N)}(\omega'_N) + \frac{1}{m}\right)\right]\right) \\ &= \frac{p^{(N)}(\omega_N) + \frac{1}{m}}{p^{(N)}(\omega'_N) + \frac{1}{m}} \stackrel{m\to\infty}{\longrightarrow} \frac{p^{(N)}(\omega_N)}{p^{(N)}(\omega'_N)} \end{aligned}$$

in accordance with (4.3).

On the other hand if $\omega' \in \mathrm{supp}(p)$ but $\omega \notin \mathrm{supp}(p)$, then there is an $i \in \{1, \dots, N-1\}$ with $\omega_N \neq \pi_i^{-1}(\omega_i)$ or $p^{(N)}(\omega_N) = 0$. In both cases

$$\lim_{m\to\infty} \frac{q^{(m)}(\omega)}{q^{(m)}(\omega')} = 0,$$

again in accordance with (4.3). As the $p^{(m)}$ are probability distributions, we have shown that $\lim_{m\to\infty} q^{(m)} = p$. □

(Received September 26, 2005.)

REFERENCES

[1] E. Aarts and J. Korst: Simulated Annealing and Boltzmann Machines. Wiley, New York 1989.
[2] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski: A learning algorithm for Boltzmann machines. Cognitive Science *9* (1985), 147–169.
[3] M. Aigner: Combinatorial Theory, Classics in Mathematics. Springer–Verlag, Berlin 1997.
[4] S. Amari: Information geometry on hierarchy of probability distributions. IEEE Trans. Inform. Theory *47* (2001), 1701–1711.
[5] S. Amari, K. Kurata, and H. Nagaoka: Information geometry of Boltzmann machines. IEEE Trans. Neural Networks *3* (1992), 2, 260–271.
[6] N. Ay: An information-geometric approach to a theory of pragmatic structuring. Ann. Probab. *30* (2002), 416–436.
[7] N. Ay: Locality of global stochastic interaction in directed acyclic networks. Neural Computation *14* (2002), 2959–2980.
[8] R. Linsker: Self-organization in a perceptual network. IEEE Computer *21* (1988), 105–117.
[9] F. Matúš and N. Ay: On maximization of the information divergence from an exponential family. In: Proc. WUPES'03 (J. Vejnarová, ed.), University of Economics, Prague 2003, pp. 199–204.
[10] C. E. Shannon: A mathematical theory of communication. Bell System Tech. J. *27* (1948), 379–423, 623–656.
[11] G. Tononi, O. Sporns, and G. M. Edelman: A measure for brain complexity: Relating functional segregation and integration in the nervous system. Proc. Nat. Acad. Sci. U. S. A. *91* (1994), 5033–5037.

Nihat Ay, Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D 04103 Leipzig. Germany.
e-mail: nay@mis.mpg.de

Andreas Knauf, Mathematisches Institut, Friedrich–Alexander–Universität Erlangen–Nürnberg, Bismarckstraße 1 1/2, D-91054 Erlangen. Germany.
e-mail: knauf@mi.uni-erlangen.de