Max-Planck-Institut
für Mathematik
in den Naturwissenschaften
Leipzig

Maximizing Multi-Information

by

Nihat Ay and Andreas Knauf

Preprint no.: 42 2003

MAXIMIZING MULTI-INFORMATION

NIHAT AY AND ANDREAS KNAUF

ABSTRACT. We investigate the structure of the global maximizers of stochastic interdependence, which is measured by the Kullback-Leibler divergence of the underlying joint probability distribution from the exponential family of factorizable random fields (multi-information). As a consequence of our structure results, it comes out that random fields with globally maximal multi-information are contained in the topological closure of the exponential family of pair interactions.

Index Terms — Multi-information, exponential family, Kullback-Leibler divergence, pair-interaction, infomax principle, Boltzmann machine, neural networks.

1. Introduction

The starting point of this article is a geometric interpretation of the interdependence 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 $\mathcal{F}$.

Date: April 27, 2003.

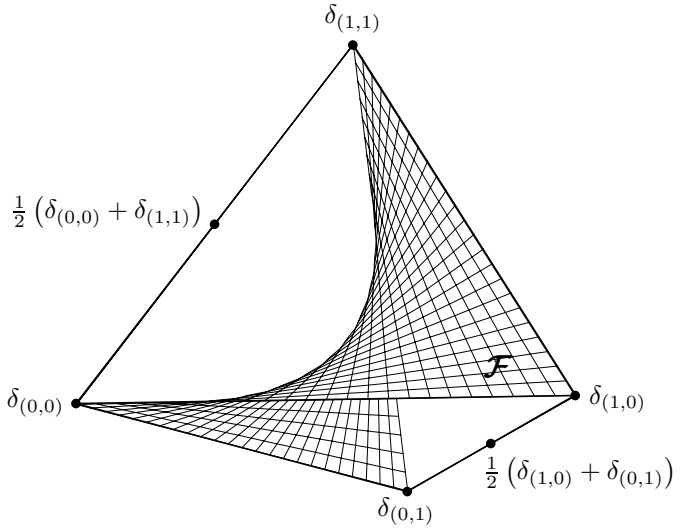

FIGURE 1: The simplex of probability distributions.

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 [Sh] as a fundamental quantity that provides a measure of the capacity of a communication channel. Motivated by so-called *Infomax principles* within the field of neural networks [Li, TSE], we have investigated maximizers of the interdependence [Ay1, Ay2] 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 our work by providing some information about 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 ([Ay1], Section 3.4, problem (ii); [Ay2], Section 4.2, Problem 4.2.3).

2. **Preliminaries**

2.1. **Notations and Definitions.** 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 $\overline{\mathcal{P}}(\Omega)$:

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

3.5 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:

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

A subset $\mathcal{E}$ of $\mathcal{P}(\Omega)$ is called *exponential family* if there exist a point $p_0 \in \mathcal{P}(\Omega)$ and a subspace $\mathcal{T}$ of $\mathbb{R}^\Omega$ such that $\mathcal{E}$ is the image of the map $\mathcal{T} \to \mathcal{P}(\Omega)$,

$$X = \big(X(\omega)\big)_{\omega\in\Omega} \quad\mapsto\quad \sum_{\omega\in\Omega} \frac{p_0(\omega)\exp\big(X(\omega)\big)}{\sum_{\omega'\in\Omega} p_0(\omega')\exp\big(X(\omega')\big)}\, e_\omega.$$

We say that $\mathcal{E}$ contains p_0 and is generated 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 (KL) divergence $D : \overline{\mathcal{P}}(\Omega) \times \overline{\mathcal{P}}(\Omega) \to \overline{\mathbb{R}}_+$,

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

to define $D_\mathcal{E} : \overline{\mathcal{P}}(\Omega) \to \mathbb{R}_+$,

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

If there exists a probability distribution $p' \in \mathcal{E}$ that satisfies

$$D_\mathcal{E}(p) \;=\; D(p \,\|\, p'),$$

then we say that p is projectable onto $\mathcal{E}$. The set of projectable elements of $\overline{\mathcal{P}}(\Omega)$ is denoted by $\operatorname{dom}\mathcal{E}$.

2.2. **Previous Results.** In this section we present some previous results from [Ay1] on maximizing the KL-divergence from an exponential family.

Theorem 2.1. *(Prop. 3.2 of* [Ay1]*) Let $\mathcal{E}$ be an exponential family in $\mathcal{P}(\Omega)$, and let p be a probability distribution in* $\operatorname{dom}\mathcal{E}$ *that locally maximizes the function $D_{\mathcal{E}}$. Then the following bound on the support of p holds:*

$$|\operatorname{supp} p| \;\leq\; \dim\mathcal{E} + 1.$$

Theorem 2.2. *(Thm. 3.5 of* [Ay1]*) Let $\mathcal{E}$ be an exponential family in $\mathcal{P}(\Omega)$ with dimension d. Then there exists an exponential family $\mathcal{E}^* \subset \mathcal{P}(\Omega)$ with dimension less than or equal to $\frac{1}{2}(d^2+7d+4)$ such that the topological closure of $\mathcal{E}^*$ contains all projectable points in $\overline{\mathcal{P}}(\Omega)$ that locally maximize the function $D_{\mathcal{E}}$.*

Theorems 2.1 and 2.2 are quite general. In this article we are interested in applications and generalizations of these statements within the setting of random fields where the set Ω has a product structure. More precisely, we consider the set $V := \{1,\ldots,N\}$ of *units*, $N \geq 2$, and corresponding sets Ω_i, $i \in V$, 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 $S \subset V$, the set of configurations on S is given by the product $\Omega_S := \times_{i\in S}\Omega_i$. The elements of $\overline{\mathcal{P}}(\Omega_S)$ are the *random fields* on S. One has the natural restriction

$$X_S : \Omega_V \to \Omega_S \quad , \quad (\omega_i)_{i\in V} \mapsto (\omega_i)_{i\in S},$$

which induces the projection

$$\overline{\mathcal{P}}(\Omega_V) \to \overline{\mathcal{P}}(\Omega_S) \quad , \quad p \mapsto p_S,$$

where p_S denotes the image measure of p under the variable X_S. For $i \in V$ we write p_i instead of $p_{\{i\}}$. A probability distribution $p \in \overline{\mathcal{P}}(\Omega_V)$ is called *factorizable* if it satisfies

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

It is well known that 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 $\mathcal{F}$. If we have a strictly positive and factorizable probability distribution p, that is $p \in \mathcal{F}$, then of course $D_{\mathcal{F}}(p) = 0$. Thus, this distance function can be interpreted as a measure that quantifies the stochastic interdependence of the units in V. The following entropic representation of $D_{\mathcal{F}}$ is well known (see [Am]):

$$I(p) \;:=\; D_{\mathcal{F}}(p) \;=\; \sum_{i=1}^{N} H_i(p) - H(p).$$

Here, the H_i's denote the marginal entropies and H is the global entropy. This measure of stochastic interdependence of the units is a generalization of the mutual information, which is called *multi-information*. The application of the Theorems 2.1 and 2.2 to the exponential family $\mathcal{F}$ leads to the following statements on local maximizers of the multi-information $I = D_{\mathcal{F}}$:

Corollary 2.3. *Let $p \in \operatorname{dom} \mathcal{F}$ be a probability distribution that locally maximizes the multi-information. Then*

$$|\operatorname{supp} p| \;\leq\; \sum_{i=1}^{N} (n_i - 1) + 1.$$

Corollary 2.4. *There exists an exponential family $\mathcal{F}^*$ with*

$$\dim \mathcal{F}^* \;\leq\; \frac{1}{2}\left(\sum_{i,j=1}^{N}(n_i-1)(n_j-1)+7\sum_{i=1}^{N}(n_i-1)+4\right)$$

that contains all local maximizers of I that are projectable onto $\mathcal{F}$ in its topological closure. In particular, in the case $N \geq 8$, and $n_i = 2$ for all i, $\dim \mathcal{F}^* \leq N^2$.

This article improves the upper bounds that appear in the Corollaries 2.3 and 2.4 by considering not the local maximizers of I but certain global maximizers (we will frequently use the term *maximizer* without any specification instead of *global maximizer*). Furthermore, we will obtain a characterization of the structure of such maximizers. The general theory is presented in the following Section 3.

3. Global Maximizers of Multi-Information – General Theory

The maximal value of I is of course bounded as

$$I(p) \;=\; \sum_{i=1}^{N} H_i(p) \;-\; H(p) \;\leq\; \sum_{i=1}^{N} \ln(n_i).$$

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

Theorem 3.1. *Let p be a probability distribution on Ω_V. Then:*

$$I(p) \;\leq\; \sum_{i=1}^{N-1} \ln(n_i). \tag{3.1}$$

In the following, the set of probability distributions that satisfy $I(p) = \sum_{i=1}^{N-1} \ln(n_i)$ is denoted by $\mathcal{M}(\Omega_1, \ldots, \Omega_N)$. In the case $\Omega_i = \{1, \ldots, n_i\}$ for all i we also use the notation $\mathcal{M}(n_1, \ldots, n_N)$. From Theorem 3.1 we know that $\mathcal{M}(\Omega_1, \ldots, \Omega_N)$

coincides with the set of maximizers of I if it is nonempty. The next theorem characterizes the probability distributions in $\mathcal{M}(\Omega_1, \ldots, \Omega_N)$.

Theorem 3.2. *Let p be a probability distribution on Ω_V. Then $p \in \mathcal{M}(\Omega_1, \ldots, \Omega_N)$ if and only if there exist a probability distribution $p^{(N)} \in \overline{\mathcal{P}}(\Omega_N)$ and surjective maps $\pi_i : \Omega_N \to \Omega_i$, $i = 1, \ldots, N-1$, with*

$$p^{(N)}\{\pi_i = \omega_i\} = \frac{1}{n_i} \quad , \quad \omega_i \in \Omega_i, \tag{3.2}$$

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

$$p(\omega_1, \ldots, \omega_N) = \begin{cases} p^{(N)}(\omega_N), & \text{if } \omega_i = \pi_i(\omega_N),\ i = 1, \ldots, N-1, \\ 0 \quad , & \text{otherwise.} \end{cases} \tag{3.3}$$

We have the following implications of Theorem 3.2:

Corollary 3.3.

(1) *$\mathcal{M}(\Omega_1, \ldots, \Omega_N) \neq \emptyset$ if the least common multiple*

$$\operatorname{lcm}(n_1, \ldots, n_{N-1}) \leq n_N. \tag{3.4}$$

For example this is the case if

- *there are only $N = 2$ units or*
- *all units are identical ($n_1 = \ldots = n_N$).*

We will consider these important cases in more detail.

(2) *Condition (3.4) is not necessary for having $\mathcal{M}(\Omega_1, \ldots, \Omega_N) \neq \emptyset$. For example $N = 3$ units with $n_2 = n_1 + 1$ and $n_3 \geq 2\, n_2$ pass this test.*

(3) *In the case $\mathcal{M}(\Omega_1, \ldots, \Omega_N) \neq \emptyset$, a probability distribution p globally maximizes the multi-information if and only if it maximizes the mutual information between i and N for all $i = 1, \ldots, N-1$.*

(4) *If n_i divides n_{i+1} for all $i = 1, \ldots, N-2$ then there exists a probability distribution p that maximizes the mutual information between i and j for all $i \neq j$.*

In this case, p is an element of $\mathcal{M}(\Omega_1, \ldots, \Omega_N)$, and therefore it also maximizes the multi-information.

(5) *In the case $n_{N-1} = n_N$, the set $\mathcal{M}(\Omega_1, \ldots, \Omega_N)$ is non-empty if and only if each n_i divides n_N.*

Examples 3.4.

(1) The set $\mathcal{M}(2, 3)$ is non-empty. In Section 4.1, we will prove that this set is homeomorphic to S^1.

(2) Consider identical units $1, \ldots, N$ with $\Omega_i = \{1, \ldots, n\}$ for all i, and partition the set $V = \{1, \ldots, N\}$ into non-empty subsets $V_1, \ldots, V_r$, with $V = V_1 \uplus \cdots \uplus V_r$, $|V_1| \leq \cdots \leq |V_r|$. Then by Corollary 3.3 (4) $\mathcal{M}(\Omega_{V_1}, \ldots, \Omega_{V_r}) \neq \emptyset$. Furthermore, each probability distribution that maximizes the mutual informations of V_j and V_{j+1}, for all $j = 1, \ldots, r-1$, also maximizes the multi-information of the parts $V_1, \ldots, V_r$. The maximization of the mutual informations of V_j and V_{j+1} is related to Linsker's [Li] *Infomax principle*, which is known in the field of neural networks as a first principle for the learning in the early visual system of mammals. In Linsker's model, the V_j's are given by the layers of a feed-forward network.

(3) By Corollary 3.3 (1) resp. (2) $\mathcal{M}\big(2, 4, 4\big)$ and $\mathcal{M}(2, 3, 4)$ are non-empty whereas by (5) $\mathcal{M}\big(3, 4, 4\big)$ is empty.

In the Section 4.1 we will discuss the two situations of Corollary 3.3 (1) more precisely. In this section, without specifying the configuration sets Ω_i we assume $\mathcal{M}(\Omega_1, \ldots, \Omega_N) \neq \emptyset$ and derive some refinements of the Corollaries 2.3 and 2.4 for the global maximizers of stochastic interdependence. First we note that the support of a probability distribution p with the structure that is described in Theorem 3.2 is bounded as

$$n_{N-1} \;\leq\; |\mathrm{supp}\, p| \;\leq\; n_N.$$

This reduces the upper bound $n_N + \sum_{i=1}^{N-1}(n_i - 1)$ from Corollary 2.3. In the following, we explicitly define an exponential family that contains all elements of $\mathcal{M}(\Omega_1, \ldots, \Omega_N)$ in its topological closure. The main idea behind our construction is to interpret the probability distributions with the structure that is described in Theorem 3.2 as the limits of the Gibbs distributions

$$p^{(m)}(\omega) \;:=\; \frac{\exp E^{(m)}(\omega)}{\sum_{\omega' \in \Omega} \exp E^{(m)}(\omega')},$$

with the energies $E^{(m)} \in \mathbb{R}^{\Omega_V}$,

$$E^{(m)}(\omega) \;:=\; \ln\left(p^{(N)}(\omega_N) + 1/m\right) + \;\; m \sum_{i=1}^{N-1} \delta_{\omega_i, \pi_i(\omega_N)}. \tag{3.5}$$

One obtains $p = \lim_{m\to\infty} p^{(m)}$ as

$$\left(\lim_{m\to\infty} p^{(m)}\right)(\omega_1, \ldots, \omega_N) \;=\; p^{(N)}(\omega_N) \prod_{i=1}^{N-1} \delta_{\omega_i, \pi_i(\omega_N)}.$$

We are now going to show that $\mathcal{M}(\Omega_1, \ldots, \Omega_N)$ lies in the closure of a low dimensional exponential family $\mathcal{F}^*$ containing the center $c(\omega) = \frac{1}{|\Omega_V|}$, $\omega \in \Omega_V$, of $\mathcal{P}(\Omega_V)$. But first we discuss the local geometry in the "tangent space" at c. Given a subset $S \subset V = \{1, \ldots, N\}$, we write $\omega \in \Omega_V$ in the form $\omega \;=\; (\omega_S, \omega_{V\setminus S})$ with $\omega_S \in \Omega_S$, $\omega_{V\setminus S} \in \Omega_{V\setminus S}$. The *conditional expectation* $\mathbb{E}_S \in \mathrm{End}(\mathbb{R}^{\Omega_V})$ with respect to c is then given by

$$(\mathbb{E}_S f)(\omega_S, \omega_{V\setminus S}) \;:=\; \frac{1}{|\Omega_{V\setminus S}|} \sum_{\omega'_{V\setminus S} \in \Omega_{V\setminus S}} f(\omega_S, \omega'_{V\setminus S}).$$

$\mathbb{E}_S$ is the orthogonal projection onto the $|\Omega_S|$–dimensional subspace $\mathcal{I}_S$ of functions on Ω_V that are X_S-measurable.

In a statistical mechanics interpretation, for each $k \in \{0, 1, \ldots, N\}$ the functions in the subspace

$$\mathcal{I}^{(k)} \;:=\; \sum_{|S|=k} \mathcal{I}_S$$

are interactions of order at most k, and the inclusions

$$\mathbb{R} \cong \mathcal{I}^{(0)} \subset \ldots \subset \mathcal{I}^{(k-1)} \subset \mathcal{I}^{(k)} \subset \ldots \subset \mathcal{I}^{(N)} = \mathbb{R}^{\Omega_V}$$

allow us to decompose the function space orthogonally into the direct sum

$$\mathbb{R}^{\Omega_V} = \bigoplus_{k=0}^{N} \widetilde{\mathcal{I}}^{(k)},$$

$\widetilde{\mathcal{I}}^{(k)} \subset \mathcal{I}^{(k)}$ denoting the orthogonal complement of $\mathcal{I}^{(k-1)}$ (with $\mathcal{I}^{(-1)} := \{0\}$). We let $\mathcal{F}^{(k)}$ be the exponential family through c generated by $\mathcal{I}^{(k)}$, and similarly $\widetilde{\mathcal{F}}^{(k)}$ the one generated by $\widetilde{\mathcal{I}}^{(k)}$. Identifying the units Ω_i with the abelian groups $\mathbb{Z}_{n_i}$, we obtain a basis of $\widetilde{I}^{(k)}$ (depending on the bijections $\Omega_i \to \mathbb{Z}_{n_i}$) by taking all products of characters of the units of which exactly k are nontrivial.
In this terminology, we have the following hierarchy of exponential families:

$$\mathcal{F}^{(0)} \subset \mathcal{F}^{(1)} \subset \cdots \subset \mathcal{F}^{(N)},$$

where

$$\mathcal{F}^{(0)} = \{c\}, \qquad \mathcal{F}^{(1)} = \mathcal{F}, \qquad \mathcal{F}^{(N)} = \mathcal{P}(\Omega).$$

The multi-information vanishes exactly on $\mathcal{F}^{(1)}$. We are interested in the lowest order k such that the set $\mathcal{M}(\Omega_1, \ldots, \Omega_N)$ is contained in the topological closure of $\mathcal{F}^{(k)}$. Of course, the first possible candidate for this is given by $k = 2$. On the other hand, from Theorem 3.5 below it immediately follows that $k = 2$ is also sufficient.

In order to obtain all probability distributions $p \in \mathcal{M}(\Omega_1, \ldots, \Omega_N)$ as limit points of one exponential family, we introduce for $i = 1, \ldots, N-1$ the subspaces

$$\mathrm{M}_i := \left\{ F \in \mathrm{Lin}(\mathbb{R}^{\Omega_i}, \mathbb{R}^{\Omega_N}) \,:\, F(\mathbb{1}_i) = 0, \text{ and if } n_i = n_N \text{ then } F^{\mathrm{ad}}(\mathbb{1}_N) = 0 \right\},$$

where $\mathbb{1}_i \in \mathbb{R}^{\Omega_i}$ is the vector with unit entries, and F^{ad} is the adjoint of F with respect to $\langle \cdot, \cdot \rangle$. We have

$$\dim \mathrm{M}_i = (n_N - \delta_{n_i, n_N})(n_i - 1).$$

Then for $F := (F_1, \dots, F_{N-1}) \in \bigoplus_{i=1}^{N-1} \mathrm{M}_i =: \mathrm{M}$ we define the energy

$$E^F : \Omega \to \mathbb{R}, \qquad E^F(\omega) = \sum_{i=1}^{N-1} \langle e_{\omega_N}, F_i(e_{\omega_i}) \rangle \,. \tag{3.6}$$

Obviously E^F is in the space $\mathcal{I}^{(2)}$ of pair interactions. The linear map

$$E : \mathrm{M} \to \mathbb{R}^{\Omega_V}, \qquad F \mapsto E^F$$

has rank

$$\operatorname{rank} E \;=\; \dim \mathrm{M} \;=\; \sum_{i=1}^{N-1} (n_N - \delta_{n_i,n_N})(n_i - 1).$$

More precisely, we have the following theorem:

Theorem 3.5. *Let $\mathcal{F}^*$ be the exponential family containing the center c of $\mathcal{P}(\Omega_V)$ and generated by the image space of E. Then*

$$\dim \mathcal{F}^* \;=\; \operatorname{rank} E, \qquad \textit{and} \qquad \mathcal{M}(\Omega_1, \dots, \Omega_N) \;\subset\; \overline{\mathcal{F}^*}.$$

Furthermore, $\mathcal{F}^* \subset \mathcal{F}^{(2)}$.

Remarks 3.6.

(1) Note that for fixed n_N, the dimension of $\mathcal{F}^*$ increases linearly in $\dim \mathcal{F}$. This improves the quadratic dimension bound in Corollary 2.4.

(2) For N equal units, $N \geq 2$, we have $\mathcal{F}^* \subset \widetilde{\mathcal{F}}^{(2)}$, and

$$\dim \mathcal{F}^* \;=\; (N-1)(n-1)^2 \;\leq\; \frac{1}{2} N(N-1)(n-1)^2 \;=\; \dim \left(\widetilde{\mathcal{F}}^{(2)} \right). \tag{3.7}$$

Note that the difference between the right-hand side and the left-hand side of inequality (3.7) increases quadratically in N.

(3) For each $F \in \mathrm{M}$, the energy (3.6) is induced from the potential $U_S^F : \Omega_V \to \mathbb{R}$, $S \subset V$, given by $U_{\{i,N\}}^F(\omega) = \langle e_{\omega_N}, F_i(e_{\omega_i}) \rangle$, $i = 1, \dots, N-1$, and $U_S \equiv 0$ if

$S \notin \{\{i, N\} \ : \ i = 1, \ldots, N-1\}$. This potential is a neighbour potential with respect to the following star-like spanning tree on the vertex set V:

$$\partial(N) = \{1, \ldots, N-1\}, \qquad \partial(i) = \{N\}, \ \ i = 1, \ldots, N-1. \tag{3.8}$$

Theorem 3.5 implies that the set $\mathcal{M}(\Omega_1, \ldots, \Omega_N)$ is in the topological closure of the exponential family of Gibbs distributions relative to the neighborhood system (3.8). But obviously, in the case of N equal units, for any other spanning tree on V, the exponential family of Gibbs distributions relative to the corresponding neighborhood system also contains $\mathcal{M}(\Omega_1, \ldots, \Omega_N)$ in its topological closure.

4. **Examples**

4.1. **The Case of Two Units.** We now discuss the case of two units, i.e. $N = 2$. In this case, the set

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

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 faces. In order to do that, we consider for $\Omega_1^* := \Omega_1 \cup \{0\}$ the following set of maps

$$\mathcal{S} \ := \ \{\pi : \ \Omega_2 \ \to \ \Omega_1^* \ : \ \pi(\Omega_2) \supset \Omega_1\}. \tag{4.9}$$

The relation

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

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

Example 4.1. 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 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).

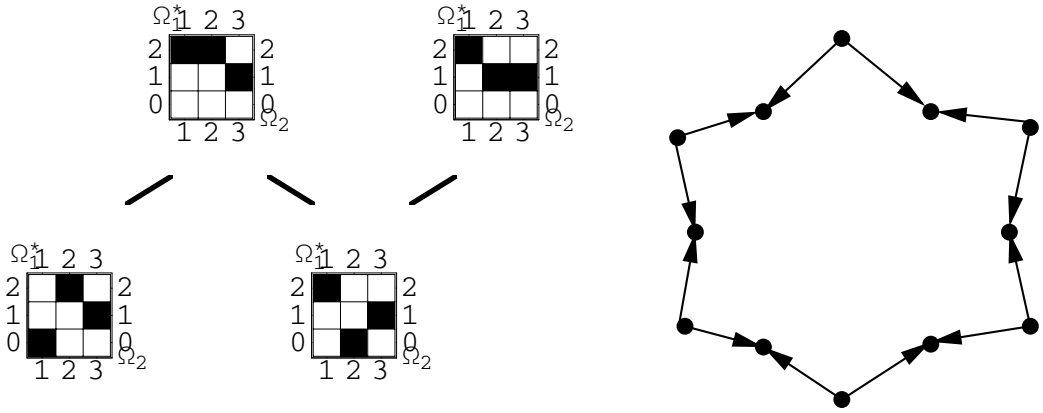

FIGURE 2: The posets for $\Omega_1 = \{1, 2\}$, $\Omega_2 = \{1, 2, 3\}$.

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

Lemma 4.2. *The poset (4.9) is connected if and only if* $n_1 < n_2$.

Given $\pi \in \mathcal{S}$ we consider the set

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

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

Theorem 4.3.

(1) *The set of global maximizers of the mutual information is a disjoint union*

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

of relatively open faces $\mathcal{M}_\pi(\Omega_1, \Omega_2)$.

(2) *These faces have dimension*

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

and there are $n_1!\binom{n_2}{l} S_{l,n_1}$ *faces* $\mathcal{M}_\pi(\Omega_1, \Omega_2)$ *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 4.4. Continuing Example 4.1, 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.

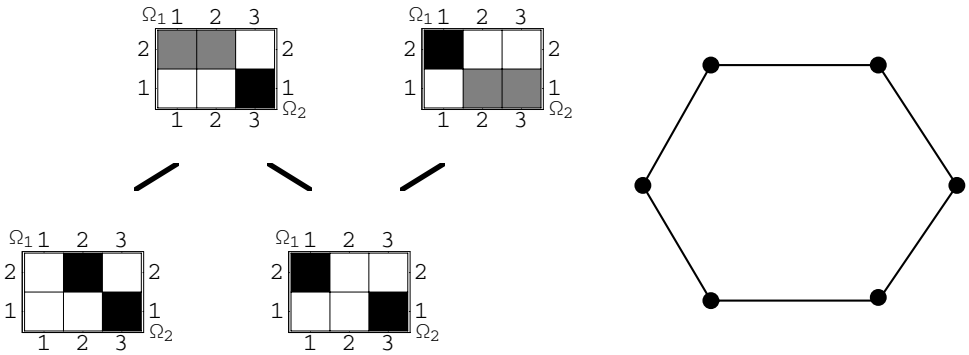

FIGURE 3: The structure of $\mathcal{M}(2, 3)$.

4.2. **The Case of N Equal Units.** This section deals with the important example of N units that have identical configuration sets: $\Omega_1 = \cdots = \Omega_N = \{0, 1, \ldots, n-1\}$. In that situation, Theorem 3.2 has the following direct implication. To simplify notation, we set

$$\mathcal{M}(N \times n) \;:=\; \mathcal{M}(\Omega_1, \ldots, \Omega_N).$$

Theorem 4.5. *The set $\mathcal{M}(N \times n)$ 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 π_i, $i = 1, \ldots, N-1$, are permutations of $\{0, 1, \ldots, n-1\}$. This implies

$$|\mathcal{M}(N \times n)| \;=\; (n!)^{N-1}, \tag{4.10}$$

and for all $p \in \mathcal{M}(N \times n)$,

$$I(p) \;=\; (N-1) \cdot \ln(n),$$

$$|\operatorname{supp} p| \;=\; n. \tag{4.11}$$

Thus according to (4.10), the number of the maximizers of the multi-information grows exponentially in N. In particular, for binary units the set $\mathcal{M}(N \times 2)$ has 2^{N-1} elements. In view of this fact, it is interesting that according to the following theorem, for all N there exists an exponential family with dimension less than or equal to 5 that approximates all 2^{N-1} elements of $\mathcal{M}(N \times n)$.

Theorem 4.6. *There exists an exponential family $\mathcal{G}^*$ of dimension less than or equal to $\frac{1}{2}(n^2 + 3n)$ with $\mathcal{M}(N \times n) \subset \overline{\mathcal{G}^*}$.*

This existence theorem should be compared with the explicitly constructed exponential family $\mathcal{F}^*$ of Theorem 3.5, which has dimension $(N-1)(n-1)^2$. Theorem 4.6 provides another exponential family $\mathcal{G}^*$ of N–independent dimension. However in this article, we are guided by ideas from physics and biology insofar we prefer to use two-body interactions like the ones leading to the definition of $\mathcal{F}^*$ instead of the multi-body interactions associated with the exponential family $\mathcal{G}^*$.

Example 4.7. For $N = n = 2$, the exponential family $\mathcal{F}^*$ of Theorem 3.5, that approximates the distributions with maximal mutual information, has dimension one. In Figure 1, we obtain this family by simply taking the convex combinations of the two maximizers:

$$\mathcal{F}^* \;=\; \left\{ \frac{1-\lambda}{2}\left(\delta_{(0,0)}+\delta_{(1,1)}\right) \;+\; \frac{\lambda}{2}\left(\delta_{(1,0)}+\delta_{(0,1)}\right) \;:\; 0<\lambda<1 \right\}.$$

The tangent space of this exponential family is spanned by the characteristic function of the set $\{(0,0),(1,1)\}$.

As stated in Theorem 3.5, pair interactions of the units are sufficient for globally maximal multi-information. For N binary units with $\Omega_i = \{0,1\} \subset \mathbb{R}$, the probability distributions $p \in \mathcal{F}^{(2)}$ on $\{0,1\}^N$, which are given by pair interactions, have the following energy expansion:

(4.12)

$$\ln p(\omega_1,\ldots,\omega_N) \;=\; const + \sum_{i=1}^{N} \theta_i\,\omega_i + \sum_{\substack{i,j=1\\ i<j}}^{N} \theta_{ij}\,\omega_i\,\omega_j, \quad (\omega_1,\ldots,\omega_N) \in \{0,1\}^N.$$

In the field of neural networks, this exponential family is known as family of *Boltzmann machines* ([AHS], [AKN]). The units are interpreted as neurons that have the two configurations "0 = not active" and "1 = active". The parameter θ_{ij} is usually interpreted as the synaptic strength between the neurons i and j, and the parameter θ_i is the threshold value of neuron i. Note that this interpretation,

which requires the symmetry of θ_{ij}, is biologically not justified. Nevertheless, Boltzmann machines represent an important example of artificial neural networks that turned out to be very fruitful for the conceptual understanding of cognitive systems. Furthermore, there are many applications to discrete optimization that are not directly related to neural networks (see [AK]).
We have the following special case of the Theorem 3.5.

Corollary 4.8. *The exponential family of Boltzmann machines approximates the global maximizers of the multi-information.*

Note that the dimensionality of the Boltzmann exponential family is of order N^2, whereas according to Theorem 3.5 the $(N-1)$-dimensional subfamily $\mathcal{F}^*$ is sufficient for approximating the maximizers of the multi-information.

Remark 4.9. Not the probability distribution itself but the corresponding *Glauber dynamics* is associated with the name "machine". It is defined as the following inhomogeneous Markov chain:

- Initialization:
  - Choose an initial configuration $\omega(0)$
- Transition $t \to t+1$:
  - *Step 1*: Choose one unit $i(t)$ with probability $\frac{1}{N}$
  - *Step 2*: Compute

$$h_{i(t)}\big(\omega(t)\big) \;:=\; \sum_{j=1}^{N} \theta_{i(t)\,j}\,\omega_j(t) \;+\; \theta_{i(t)}.$$

  - *Step 3*: Set

$$\omega_{i(t)}(t+1) \;:=\; \begin{cases} 1 & \text{with probability} \quad \frac{1}{1+\exp(-h_{i(t)}(\omega(t)))} \\ 0 & \text{with probability} \quad 1-\frac{1}{1+\exp(-h_{i(t)}(\omega(t)))} \end{cases}$$

It is well known that the Glauber dynamics converges to the stationary distribution (4.12). Thus, Corollary 4.8 states that the Glauber dynamics is able to generate probability distributions arbitrarily close to the maximizers of the multi-information.

5. **Proofs**

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

In order to prove Theorem 3.1 we need the following proposition:

Proposition 5.1. *Let $V_1, \ldots, V_r$ be a partition of $\{1, \ldots, N\}$. Then*

$$I \;=\; I_{\{V_1,\ldots,V_r\}} + \sum_{j=1}^r I_{V_j}. \tag{5.1}$$

Proof. Let p be a probability distribution on Ω_V. Then

$$\begin{aligned} I(p) &= \sum_{i=1}^N H_i(p) - H(p) \\ &= \sum_{j=1}^r \left(\sum_{i \in V_j} H_i(p) - H_{V_j}(p) \right) + \sum_{j=1}^r H_{V_j}(p) - H(p) \\ &= \sum_{j=1}^r I_{V_j}(p) + I_{\{V_1,\ldots,V_r\}}(p). \end{aligned}$$

□

Proof of Theorem 3.1. In order to prove the estimate (3.1) we choose the partition $\{1\}, \{2, \ldots, N\}$ of $\{1, \ldots, N\}$ and apply equation (5.1):

$$\begin{aligned} I &= I_{\{\{1\},\{2,\ldots,N\}\}} + I_{\{2,\ldots,N\}} \\ &= H_1 - H_{(1\,|\,2,\ldots,N)} + I_{\{2,\ldots,N\}} \\ &\leq H_1 + I_{\{2,\ldots,N\}} \\ &\leq \ln(n_1) + I_{\{2,\ldots,N\}}. \end{aligned}$$

Iterating this procedure implies

$$\begin{aligned} I &= \sum_{i=1}^{N-2} \ln(n_i) + I_{\{N-1,N\}} \\ &\leq \sum_{i=1}^{N-2} \ln(n_i) + \big(H_{N-1} - H_{(N-1|N)}\big) \\ &\leq \sum_{i=1}^{N-2} \ln(n_i) + \ln(n_{N-1}). \end{aligned}$$

□

Proof of Theorem 3.2. If a probability distribution p on Ω_V has the form (3.3) with a distribution $p^{(N)} \in \overline{\mathcal{P}}(\Omega_N)$ and surjective maps $\pi_i : \Omega_N \to \Omega_i$ that satisfy (3.2), then $I(p) = \sum_{i=1}^{N-1} \ln(n_i)$:

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

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

$$H_i(p) \;=\; \ln(n_i), \qquad i = 1, \ldots, N-1. \tag{5.2}$$

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

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

From (5.2) we have

$$H(p) \;=\; \sum_{i=1}^{N} H_i(p) - I(p) \;=\; \left(\sum_{i=1}^{N-1} \ln(n_i) + H_N(p)\right) - \sum_{i=1}^{N-1} \ln(n_i) \;=\; H_N(p). \tag{5.3}$$

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

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

In these definitions we get

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

From (5.3) 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

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

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

Proof of Corollary 3.3.

(1) W.l.o.g. we set $\Omega_i := \{0, 1, \ldots, n_i - 1\}$ and use the surjections $\pi_i : \Omega_N \to \Omega_i$,

$\omega_N \mapsto \omega_N (\mathrm{mod}\, n_i)$. Defining the probability distribution p by (3.3) with

$$p^{(N)}(\omega_N) := \begin{cases} \frac{1}{L}, & \text{if } \omega_N < L \\ 0\,, & \text{if } \omega_N \geq L \end{cases} \quad \text{and} \quad L := \mathrm{lcm}\,(n_1, \ldots, n_{N-1})$$

shows $p \in \mathcal{M}(\Omega_1, \ldots, \Omega_N)$.

(2) Set $\pi_1(\omega_N) := \omega_N (\mathrm{mod}\, n_1)$ and

$$\pi_2(\omega_N) := \begin{cases} \omega_N, & \text{if } 0 \leq \omega_N \leq n_1 \\ n_1\,, & \text{if } n \leq \omega_N \leq n_3. \end{cases}$$

We use p given by (3.3) with

$$p^{(N)}(\omega_N) := \begin{cases} \frac{1}{n_1+1}\,, & \text{if } 0 \leq \omega_N < n_1 \\ \frac{1}{n_1^2+n_1}, & \text{if } n_1 \leq \omega_N < 2n_1 \\ 0 \quad , & \text{if } 2n_1 \leq \omega_N. \end{cases}$$

(3) This is an immediate consequence of (3.3) as $p \in \overline{\mathcal{P}}(\Omega_1, \ldots, \Omega_N)$ is of the form

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

if for all $j = 1, \ldots, N-1$ the image measures w.r.t. the projections $\Omega_1 \times \ldots \times \Omega_N \to \Omega_j \times \Omega_N$ have distributions $p^{(N)}(\omega_N)\delta_{\omega_j, \pi_j(\omega_N)}$.

(4) For the distribution p defined in (1) the image measure w.r.t. to the projection

$$\Omega_1 \times \ldots \times \Omega_N \to \Omega_i \times \Omega_j$$

with $1 \leq i < j \leq N-1$ has distribution

$$(\omega_i, \omega_j) \mapsto \begin{cases} \frac{1}{n_j}, & \text{if } \omega_i = \omega_j (\mathrm{mod}\, n_i) \\ 0\,, & \text{otherwise} \end{cases}$$

and thus mutual information $\ln(n_i)$.

(5) If $n_{N-1} = n_N$, π_{N-1} is a bijection so that

$$p^{(N)}(\omega_N) = \frac{1}{n_N} \quad \text{and} \quad p^{(N)}(\pi_i^{-1}(\omega_i)) = \frac{|\pi_i^{-1}(\omega_i)|}{n_N}.$$

Thus (3.2) implies $|\pi_i^{-1}(\omega_i)| = \frac{n_N}{n_i}$. □

Proof of Theorem 3.5. We show that $p \in \mathcal{M}(\Omega_1, \dots, \Omega_N)$ is the limit point of the sequence

$$p^{(m)} := \sum_{\omega\in\Omega} \frac{\exp\left(E^{F^{(m)}}(\omega)\right)}{\sum_{\omega'\in\Omega}\exp\left(E^{F^{(m)}}(\omega')\right)}\, e_\omega, \qquad m \in \mathbb{N},$$

of probability distributions $p^{(m)} \in \mathcal{F}^* \subset \mathcal{P}(\Omega_V)$, with $F^{(m)} = (F_1^{(m)}, \dots, F_{N-1}^{(m)}) \in \mathrm{M}$. Setting $\Psi_i \in \mathrm{End}\,(\mathbb{R}^{\Omega_i})$, $\Psi_i(e_{\omega_i}) := e_{\omega_i} - \frac{1}{n_i-1}\sum_{\omega_i'\neq\omega_i} e_{\omega_i'}$, we have $\Psi_i(\mathbb{1}_i) = 0$, so that in the case $n_i < n_N$

$$\widetilde{F}_i \circ \Psi_i \in \mathrm{M}_i \quad \text{if} \quad \widetilde{F}_i \in \mathrm{Lin}(\mathbb{R}^{\Omega_i}, \mathbb{R}^{\Omega_N}).$$

Similarly in the case $n_i = n_N$

$$\Psi_N \circ \widetilde{F}_i \circ \Psi_i \in \mathrm{M}_i \quad \text{if} \quad \widetilde{F}_i \in \mathrm{Lin}(\mathbb{R}^{\Omega_i}, \mathbb{R}^{\Omega_N}).$$

We set $F_i^{(m)} := \widetilde{F}_i^{(m)} \circ \Psi_i$ if $n_i < n_N$ and $F_i^{(m)} := \Psi_N \circ \widetilde{F}_i^{(m)} \circ \Psi_i$ if $n_i = n_N$ with

$$\widetilde{F}_i^{(m)}(e_{\omega_i}) \;:=\; m \sum_{\omega_N\in\pi_i^{-1}(\omega_i)} e_{\omega_N} \qquad (i = 1, \dots, N-2)$$

and

$$\widetilde{F}_{N-1}^{(m)}(e_{\omega_{N-1}}) \;:=\; \sum_{\omega_N\in\pi_{N-1}^{-1}(\omega_{N-1})} \left(m + \ln\left(p^{(N)}(\omega_N) + \frac{1}{m}\right)\right) e_{\omega_N}.$$

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

$$\begin{aligned} &\left\langle e_{\omega_N}, \widetilde{F}_i^{(m)}(e_{\omega_i})\right\rangle - \left\langle e_{\omega_N'}, \widetilde{F}_i^{(m)}(e_{\omega_i'})\right\rangle \\ =\;& \left\langle e_{\omega_N}, F_i^{(m)}(e_{\omega_i})\right\rangle - \left\langle e_{\omega_N'}, F_i^{(m)}(e_{\omega_i'})\right\rangle \end{aligned}$$

since

$$\sum_{\widetilde{\omega}_i'\neq\omega_i'} \mathbb{1}_{\pi_i^{-1}(\widetilde{\omega}_i')}(\omega_N') - \sum_{\widetilde{\omega}_i\neq\omega_i} \mathbb{1}_{\pi_i^{-1}(\widetilde{\omega}_i)}(\omega_N) \;=\; \mathbb{1}_{\pi_i^{-1}(\omega_i)}(\omega_N) - \mathbb{1}_{\pi_i^{-1}(\omega_i')}(\omega_N') \;=\; 0.$$

So

$$\frac{p^{(m)}(\omega)}{p^{(m)}(\omega')} = \exp\left(E^{F^{(m)}}(\omega) - E^{F^{(m)}}(\omega')\right) = \exp\left(E^{\widetilde{F}^{(m)}}(\omega) - E^{\widetilde{F}^{(m)}}(\omega')\right)$$
$$= \exp\left(\left[m(N-1) + \ln\left(p^{(N)}(\omega_N) + \frac{1}{m}\right)\right] - \left[m(N-1) + \ln\left(p^{(N)}(\omega'_N) + \frac{1}{m}\right)\right]\right)$$
$$= \frac{p^{(N)}(\omega_N) + \frac{1}{m}}{p^{(N)}(\omega'_N) + \frac{1}{m}},$$

in accordance with (3.3).

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

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

again in accordance with (3.3). As the $p^{(m)}$ are probability distributions, we have shown that $p^{(m)} \to p$. □

Proof of Lemma 4.2. 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

$$\sigma \in \mathcal{S}, \qquad \sigma(\omega_2) := \begin{cases} \pi(\omega_2), & \text{if } \omega_2 \neq \omega'_2 \\ 0 \quad , & \text{if } \omega_2 = \omega'_2 \end{cases}.$$

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

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

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

$$\pi'' \circ (\pi')^{-1} \mid_{\Omega_1} \colon \Omega_1 \to \Omega_1$$

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

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

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 4.3. 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 3.2 hold 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 3.2 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 3.2 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 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 [Ai], 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 the face $\mathcal{M}_\pi$ 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 Theorem 4.5. All statements directly follow from Theorem 3.2 . □

Proof of Theorem 4.6. We choose a map $\phi = (\phi_1, \ldots, \phi_n) : \Omega \to \mathbb{R}^n$ such that the points $\phi(\omega)$, $\omega \in \Omega$, are in general position; that is, each k elements of $\phi(\Omega)$ with $k \leq n+1$ are affinely independent. This property guarantees that for each set $\Sigma \subset \Omega$, $|\Sigma| = n$, there exist real numbers $a_1, \ldots, a_n, b$ such that

$$\left\{ \omega \in \Omega \ : \ \sum_{i=1}^{n} a_i \, \phi_i(\omega) = b \right\} \ = \ \Sigma \tag{5.4}$$

holds. We consider the exponential family $\mathcal{G}^*$ that is generated by c and

$$\phi_1, \ldots, \phi_n \quad , \quad \phi_i \, \phi_j \quad , \quad 1 \leq i \leq j \leq n.$$

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 4.5 we know that $|\operatorname{supp} p| = n$. 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 $a_1, \ldots, a_n, b$ satisfying (5.4) with $\Sigma = \operatorname{supp} p$. Then with

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

the sequence

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

converges to p. □

References

[Ai] M. Aigner: *Combinatorial Theory, Classics in Mathematics*, Springer, Berlin 1997

[AHS] D. H. Ackley; G. E. Hinton; T. J. Sejnowski: *A learning algorithm for Boltzmann machines*, Cognitive Science **9**, 147–169 (1985)

[AK] E. Aarts; J. Korst: *Simulated Annealing and Boltzmann Machines*, Wiley, New York 1989

[AKN] S. Amari; K. Kurata; H. Nagaoka: *Information Geometry of Boltzmann Machines*, IEEE Trans. NN. **3**, No. 2, 260–271 (1992)

[Am] S. Amari: *Information geometry on hierarchy of probability distributions*, IEEE Trans. IT **47**, 1701–1711 (2001)

[Ay1] N. Ay: *An Information-Geometric Approach to a Theory of Pragmatic Structuring*, Ann. Prob. **30**, 416–436 (2002)

[Ay2] N. Ay: *Locality of Global Stochastic Interaction in Directed Acyclic Networks*, Neural Computation **14**, 2959–2980 (2002)

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

[Sh] C. E. Shannon: *A mathematical theory of communication*, Bell System Tech. J. **27**, 379–423, 623–656 (1948)

[TSE] G. Tononi; O. Sporns; G. M. Edelman: *A measure for brain complexity: Relating functional segregation and integration in the nervous system*, Proc. Natl. Acad. Sci. USA **91**, 5033–5037 (1994)

Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22–26, D-04103 Leipzig, Germany

E-mail address: nay@mis.mpg.de

Mathematisches Institut, Friedrich-Alexander-Universität Erlangen-Nürnberg, Bismarckstr. 1 1/2, D-91054 Erlangen, Germany

E-mail address: knauf@mi.uni-erlangen.de