Max-Planck-Institut für Mathematik in den Naturwissenschaften Leipzig

Maximizing the divergence from a hierarchical model of quantum states

by

Stephan Weis, Andreas Knauf, Nihat Ay, and Ming-Jing Zhao

Preprint no.: 58 2014

MAXIMIZING THE DIVERGENCE FROM A HIERARCHICAL MODEL OF QUANTUM STATES

STEPHAN WEIS[1], ANDREAS KNAUF[2], NIHAT AY[1,3,4] AND MING-JING ZHAO[5]

Abstract. We quantify higher-order correlations in a composite quantum system in terms of the divergence from a family of Gibbs states with well-specified interaction patterns, known as hierarchical model. We begin with a review of factoring in a classical hierarchical model. This is just one aspect of our critical discussion of the divergence from a hierarchical model of quantum states. Then we turn to maximizers of the divergence from a Gibbs family. For example, we consider an upper bound on the support size of a local maximizer of higher-order correlation and we improve it to the square root, asymptotically for identical units. We compute the global maximizers of the mutual information of two separable qubits.

Index Terms – mutual information, multi-information, higher-order correlations, hierarchical model, factoring, Gibbs family, maximum entropy, maximizers of correlation, separable state.

AMS Subject Classification: 94A17, 81P45, 62F30.

1. Introduction

We quantify correlations in a system by measuring a distance from a set of system states whose correlations are conceived as structured in a simple way. This approach is very general and the oldest example to our best knowledge was proposed in 1997 by Vedral et al. [38] in order to have a numerical quantification of entanglement which is the quantum physical attribute responsible for phenomena such as non-locality. Here the distance is the Umegaki relative entropy, which we call *divergence*. The state of a quantum system is described by a density matrix that is a self-adjoint matrix with trace one. The divergence of the state ρ from the state σ is defined by $D(\rho\|\sigma) := \mathrm{tr}\rho(\log\rho - \log\sigma)$ if the support of ρ is included in the support of σ, and $D(\rho\|\sigma) = +\infty$ otherwise (the support of a density matrix is its image). Here and in the sequel we use the logarithm to the base 2. The divergence is a distance-like functional of pairs of quantum states, which has non-negative values and which is zero only for equal arguments. It is not a metric, though. Vedral et al. consider the divergence from the set of separable states, that is non-entangled states. This divergence is known as *relative entropy of entanglement* and has been refined to more subtle measures of entanglement, see Horodecki et al. and Modi et al. [20, 29].

Entanglement is not in the main focus of the present article. We regard as structured in a simple way the class of hierarchical models of Gibbs states $e^h/\mathrm{tr}(e^h)$ of Hamiltonians h acting on specified subsets of units. The main example is the set $\mathcal{E}^k$ of Gibbs states of Hamiltonians of maximal order k acting locally on single units of the system ($k = 1$), on pairs of units ($k = 2$), on triples of units, and so on. The Gibbs states for $k = 2$ are known as quantum Boltzmann machines [44]. We are interested in the separation of total correlation based on hierarchies of interaction

Date: June 3, 2014.

as shown in the following inclusion hierarchy:

$$\mathcal{E}^1 \subset \mathcal{E}^2 \subset \mathcal{E}^3 \subset \cdots .$$

This approach to higher-order correlations has appeared in 2001 in the setting of probability distributions in the work of Ay [4] and Amari [1], where the distance is the Kullback-Leibler divergence. Maximizers of the divergence have been studied in this setting by Ay and Knauf [6] and Rauh [34]. Weighted sums of divergences from several families $\mathcal{E}^i$ were analyzed by Ay et al. [7] because some popular correlation measures used in science are of this form. Here we distinguish classical systems, governed by a joint probability distribution, from quantum systems, whose state is described by a density matrix. Throughout, we restrict to finite-level systems, which are represented as a subalgebra of an algebra of finite-sized complex square matrices. A classical-quantum distinction will be made where results depend on the choice of the algebra, the classical case is tacitly included in the quantum case in terms of the algebra of diagonal matrices.

In the quantum setting, higher-order correlations of a quantum state ρ have been quantified in a similar approach by Niekamp et al. [30] in terms of

$$\mathcal{I}^k(\rho) := \inf\{D(\rho\|\sigma) \mid \sigma \in \mathcal{E}^k\}. \tag{1}$$

The functional $\mathcal{I}^k$ quantifies the correlations, which are not generated by correlations between k-tuples of units. Although this view appears to be common sense in statistical mechanics, we point out in the paragraph of (15) that this view is not strictly rigorous.

To this end, the first out of two focal themes of this article is a critical discussion of the functionals $\mathcal{I}^k$. One topic in this discussion is factoring the other is a maximum-entropy characterization. We summarize these ideas now. For $k = 1$ and a system of $N \in \mathbb{N}$ units, indexed by $[N] := \{1, \ldots, N\}$, the order-1 Gibbs states are tensor products $\rho_1 \otimes \cdots \otimes \rho_N$ of invertible density matrices. We will show in Section 5 that the divergence $\mathcal{I}^1(\rho)$ of a state ρ is *always* equal to the *multi-information*, that is

$$\mathcal{I}^1(\rho) = \sum_{i\in[N]} H(\rho^i) - H(\rho).$$

Here $H(\sigma) := -\operatorname{tr}\sigma\log(\sigma)$ is the *von Neumann entropy* of a density matrix σ, the density matrix ρ^i is the i-th marginal of ρ, $i \in [N]$. The multi-information $\mathcal{I}^1(\rho)$ measures the number of random bits needed to erase all correlations between the units, see Groisman et al. [17]. Therefore $\mathcal{I}^1(\rho)$ can be interpreted in terms of classical bits also in the quantum case. We obtain $\mathcal{I}^1(\rho) = D(\rho\|\rho^1 \otimes \cdots \otimes \rho^N)$ from a simple computation, hence $\mathcal{I}^1(\rho) = 0$ if and only if $\rho = \rho^1 \otimes \cdots \otimes \rho^N$. In this sense $\mathcal{I}^1$ completely characterizes factoring into product states.

Already classically, for $k \geq 2$ an analogous factoring property of Gibbs states of maximal order k, made precise in (5), is not preserved by limits. Kahle et al. [24, 23] have pointed out that there are distributions in the closure of $\mathcal{E}^k$, hence of zero divergence $\mathcal{I}^k$, which do not factor. For completeness we revisit and extend this discussion, following Geiger et al. [16]. In the quantum setting a factoring analogous to (5) is unknown for $k \geq 2$.

While factoring does not characterize states with $\mathcal{I}^k(\rho) = 0$ we want to point out that the functionals $\mathcal{I}^k$ are fully compatible with the maximum-entropy principle by von Neumann [39] and Jaynes [21]. This was known for the bulk of states ρ with $\mathcal{I}^k(\rho) = 0$, namely for invertible ρ, and Weis [42] has extended this result to

all states by showing that $\mathcal{I}^k(\rho) = 0$ if and only if ρ is a maximum-entropy state, possibly of reduced support.

Before turning to applications we mention two theoretical problems. While the multi-information $\mathcal{I}^1$ is continuous as a sum of continuous von Neumann entropies, the continuity of $\mathcal{I}^k$ needs a proof for $k \geq 2$ in a non-commutative algebra. The problem is that a set of maximum-entropy quantum states under linear constraints may not be norm closed. An example from [43] is recalled in Section 3. Although we expect that $\mathcal{I}^k$ is continuous for all $k \in \mathbb{N}$ this is an open problem for $k \geq 2$. The continuity will be needed for example to prove the convergence of algorithms in [30] that compute $\mathcal{I}^k$. Another theoretical question is whether the closure of $\mathcal{E}^k$ is an algebraic variety similar to the toric variety in the classical case in Section 2. This might hold although $\mathcal{E}^k$ has no known factoring property for $k \geq 2$.

The second focal theme of the article is the maximization of the divergence $\mathcal{I}^k$. This problem, motivated by the idea that natural systems tend to maximize structured correlations, was proposed by Ay [5] who has obtained interesting results about the divergence from Gibbs families of probability distributions. The latest result in this context is by Rauh [34]. The present article continues the work by Weis and Knauf [43, 42] and generalizes some of the methods in [5] to quantum states.

We find that a local maximizer of the divergence from a Gibbs family reflects two key features from the classical case:

- A local maximizer of the divergence from a Gibbs family $\mathcal{E}$ is the conditional distribution of its projection to $\mathcal{E}$;
- a local maximizer of the divergence from $\mathcal{E}$ is supported on a set of size of at most $\dim_{\mathbb{R}}(\mathcal{E}) + 1$.

The upper bound on the support size improves to $\sqrt{\dim_{\mathbb{R}}(\mathcal{E}) + 1}$ in the quantum setting because the state space of an n-level quantum system has dimension $n^2 - 1$ compared to $n-1$ which is the dimension of the probability simplex. For example, if all $N \in \mathbb{N}$ units of a composite system have the same number of $n \in \mathbb{N}$ levels, then the independence model $\mathcal{E}^1$ has dimension $N(n-1)$ in the classical case and $N(n^2-1)$ in the quantum case. Therefore, a local maximizer of the multi-information $\mathcal{I}^1$ has support at most $\mathcal{O}(N)$ respectively $\mathcal{O}(\sqrt{N})$, see (22) and (23). This exponential reduction shows how far a highly correlated system is from the complete randomness of the identity with support size n^N. In a loose analogy, if the classical bound was sharp, these bounds confirm that quantum systems are less chaotic than classical systems [10, 11].

Global maximizers seem to be less coherent in the classical-quantum comparison. The classification of global maximizers of the multi-information by Ay and Knauf [6], which holds in the classical setting, is not valid in the quantum setting due to the entanglement. This can be seen already in the simplest example: Two classical binary units have one bit of maximal mutual information (multi-information for two units) if the two units are in functional dependence while two qubits can have two bits of mutual information if the system is in a maximally entangled pure state. However, some of the methods in [6] can be used in the maximization of the multi-information of separable quantum states. A constrained version of this problem will be of interest in the context of classical correlations *à la* Groisman et al. [17].

This article is organized as follows. Section 2 discusses factoring of probability measures. Section 3 defines Gibbs families of quantum states and addresses theoretical issues of their divergence and of the functionals $\mathcal{I}^k$. Section 4 introduces hierarchical models of quantum states and calculates dimensions and linear bases. In Section 5 we prove that the divergence from the independence model is the multi-information. Finally, we turn to maximizers of $\mathcal{I}^k$ and of the divergence from a Gibbs family. Section 6 treats local maximizers and Section 7 is about global maximizers of the mutual information of separable states.

2. Factoring of probability distributions

We will discuss that classically, the set $\mathcal{E}^k$ of Gibbs states of maximal order k has several equivalent definitions, in terms of a monomial map, of factoring into a product and of an algebraic variety. The first two properties are not preserved by taking limits. This is somewhat similar to conditional independence [28] where the Hammersley-Clifford theorem shows several conditional independence statements equivalent to factoring into a product if a distributions has full support but not for reduced support.

We discuss hierarchical models of probability distributions in the wider context of log-linear models and monomial maps before noting that a probability distribution factors if it has a product form (5). This means something between belonging to the model and belonging to its (norm) closure. We give credit to the description of the closure as an algebraic variety and we discuss possible support sets, both for distributions in the closure and for distributions that factor.

Formally, a hierarchical model is a special case of a *log-affine model*, also known as *Gibbs family*. The definition is based on a finite state space $X := [m]$, $m \in \mathbb{N}$. We denote the probability simplex over X by

$$\Delta(X) := \{p \in \mathbb{R}^X \mid \forall i \in X : p(i) \geq 0, p(1) + \cdots + p(m) = 1\}.$$

The *support* of a real tuple $v = (v_i)_{i \in Y}$, indexed by some finite set Y is denoted $\operatorname{supp}(v) := \{i \mid v_i \neq 0\} \subset Y$. A log-affine model consists of all probability distributions $p \in \Delta(X)$ of full support $\operatorname{supp}(p) = X$, such that $\log(p) \in h_0 + H$ where $h_0 \in \mathbb{R}^X$ is a fixed function and $H \subset \mathbb{R}^X$ a fixed linear subspace. In the sequel we use *log-linear models* where $h_0 = 0$.

We follow Geiger et al. [16] and restrict the discussion to log-linear models with linear space spanned by vectors of natural numbers. Let $d \in \mathbb{N}$, $X = [m]$ and $A = (a_{ij}) \in {\mathbb{N}_0}^{d \times m}$ be a matrix with constant column sums $\sum_{i=1}^d a_{ij} = \sum_{i=1}^d a_{ik}$ for all $j, k \in X$. We denote columns of A by a_x, $x \in X$. A log-linear model $\mathcal{E}_A$ is defined by the row-span of A, which we parameterize in terms of $\theta \in \mathbb{R}^d$,

$$p_\theta(j) := Z(\theta)^{-1} \exp(\textstyle\sum_{i=1}^d \theta_i a_{ij}), \quad j \in X, \tag{2}$$

with normalization $Z(\theta) > 0$. The matrix A also defines a monomial map

$$\Phi_A : [0, \infty)^d \to [0, \infty)^m, \quad (t_1, \ldots, t_d) \mapsto (\textstyle\prod_{i=1}^d t_i^{a_{ij}})_{j=1}^m,$$

where we agree on $0^0 = 1$ and $0^\alpha = 0$ for $\alpha > 0$. A probability distribution $p \in \Delta(X)$ is said to *factor* according to A if p lies in the image of Φ_A, denoted by $\operatorname{image}(\Phi_A)$. It is easy to show by slightly generalizing the definition in (2), that

$$\operatorname{image}(\Phi_A) \cap \Delta(X) = \{p_\theta \mid \theta \in [-\infty, \infty)^d, \exists x \in X \, \forall i \in \operatorname{supp}(a_x) : \theta_i > -\infty\}. \tag{3}$$

Clearly (3) shows that all distributions factoring according to A belong to the (norm) closure $\overline{\mathcal{E}_A}$ of the log-linear model $\mathcal{E}_A$. The support sets of distributions factoring according to A can be computed by listing the support sets of distributions p_θ for all possible values of $\theta_i \in \{-\infty, 0\}$, $i = 1, \ldots, d$ meeting the condition in (3). These support sets $F \subset X$ are obviously characterized by $F \neq \emptyset$ and

$$\operatorname{supp}(a_j) \not\subset \bigcup_{k \in F} \operatorname{supp}(a_k), \quad \forall j \in X \setminus F.$$

A subset of X satisfying these conditions is called *feasible* with respect to A in [16]. This combinatorial characterization of support sets of factoring distributions will become more meaningful in the context of hierarchical models, where factoring is equivalent to an actual product form of a probability distribution, see (5) and the subsequent paragraph.

Notice, if H is the row-span of A, then for each feasible set F the log-linear model of $H|_F$ factors according to $(a_x)_{x \in F}$ and, embedded into $\Delta(X)$, it factors according to A. It is well-known, that the (norm) closure $\overline{\mathcal{E}_A}$ of $\mathcal{E}_A$ is a union of log-linear models of spaces $H|_G$, too, but for a larger class of support sets $G \subset X$. This was proved in great generality by Csiszár and Matúš [13].

It is important, because an open question in the quantum case, to observe that the closure $\overline{\mathcal{E}_A}$ is a variety. Geiger et al. have shown in [16], Theorem 3.2, that $\overline{\mathcal{E}_A}$ is the intersection of the probability simplex $\Delta(X)$ with a *non-negative toric variety* which is defined as the set of all vectors $(x_1, \ldots, x_m) \in [0, \infty)^m$ such that

$$x_1^{u_1} x_2^{u_2} \cdots x_m^{u_m} = x_1^{v_1} x_2^{v_2} \cdots x_m^{v_m}$$

holds for all vectors $u = (u_1, \ldots, u_m)$ and $v = (v_1, \ldots, v_m)$ of non-negative integers such that $u - v$ is in the kernel of A.

Let us progress to hierarchical models as studied by Lauritzen [28]. The definition is based on a composite system of units $[N]$, $N \in \mathbb{N}$, with local state spaces $X_i = [n_i]$, $n_i \in \mathbb{N}$, $i \in [N]$, and state space $X := \times_{i \in [N]} X_i$. The state space of $v \subset [N]$ is defined as $X_v := \times_{i \in v} X_i$ and we denote the truncation of $x = (x_i)_{i \in [N]} \in X$ by $x_v := (x_i)_{i \in v}$. The *factor space* of v is

$$F_v := \{h \in \mathbb{R}^X \mid h(x, y) = h(x, z)\ \forall x \in X_v\, \forall y, z \in X_{[N] \setminus v}\}.$$

We denote the power set of a set M by 2^M, and the collection of k-element subsets of M by $\binom{M}{k}$, $k \in \mathbb{N}_0$. Let $U \subset 2^{[N]}$ be a non-empty class of subsets of $[N]$. A *hierarchical model subspace* of U is defined by $H_U := \sum_{v \in U} F_v$. For instance, for each $k = 1, \ldots, N$ we can consider the set U^k of those subsets of $[N]$ that have maximal cardinality k, that is

$$U^k = \bigcup_{i=0}^{k} \binom{[N]}{i}. \tag{4}$$

Obviously,

$$U^1 \subset U^2 \subset \cdots \subset U^k \subset \cdots \subset U^N.$$

We define the *order* of a function $f \in \mathbb{R}^X$, also referred to as *Hamiltonian*, as the minimal number k for which $f \in H_{U^k}$. The corresponding *Gibbs state* of order k is defined by $p(x) = \frac{e^{f(x)}}{\sum_{x'} e^{f(x')}}$, $x \in X$.

Given any non-empty class $U \subset 2^{[N]}$, the log-linear model with hierarchical model subspace H_U is called the *hierarchical model* of U. For example, the set $\mathcal{E}^k$ of *Gibbs*

states of maximal order k is defined as the hierarchical model of U^k. Notice that $\mathcal{E}^k$ equals also the hierarchical model of $\binom{[N]}{k}$. This will make a difference in Section 4, where we use pure factor spaces in addition to factor spaces F_v.

Clearly, elements p of the hierarchical model of U can be written in the form

$$p(x) = \prod_{v \in U} \Psi_v(x_v), \quad x \in X \tag{5}$$

for functions $\{\Psi_v : X_v \to [0, \infty)\}_{v \in U}$. A probability distribution $p \in \Delta(X)$ of arbitrary support $\mathrm{supp}(p) \subset X$ not necessarily $\mathrm{supp}(p) = X$ that satisfies (5) is said to *factor* according to U.

A short calculation shows that factoring according to U, see (5), is the same as factoring according to a well-chosen matrix representation (3). More detailed, we take in (2) the matrix A with rows indexed by pairs (v, x), $v \in U$, $x \in X_v$ and with columns indexed by X such that

$$a_{(v,x),y} := \begin{cases} 1 & \text{if } y_v = x \\ 0 & \text{else} \end{cases}, \quad v \in U, x \in X_v, y \in X. \tag{6}$$

With this particular choice of matrix A, factoring according to A is equivalent to factoring according to U (we can identify $\Psi_v(y)$ with $\exp(\theta_{(v,y)})$ for all $v \in U, y \in X_v$, modulo (v, y)-independent normalization $Z(\theta)$). We call a subset of X *feasible* with respect to U if and only if it is feasible with respect to the matrix A. The notion of feasibility allows us to compute support sets of factoring distributions.

Let us consider two examples to see why factoring is not preserved by limits.

Example 2.1. The hierarchical model of $\binom{[3]}{2} = \{\{1, 2\}, \{2, 3\}, \{1, 3\}\}$ is the set $\mathcal{E}^2$ of Gibbs states of maximal order two and is known in the literature as the model of *no three-way interaction*. For three binary units we have $X = \{0, 1\}^3$ and the matrix A is

	$(0,0,0)$	$(0,0,1)$	$(0,1,0)$	$(0,1,1)$	$(1,0,0)$	$(1,0,1)$	$(1,1,0)$	$(1,1,1)$
$\{1,2\},(0,0)$	1	1	0	0	0	0	0	0
$\{1,2\},(0,1)$	0	0	1	1	0	0	0	0
$\{1,2\},(1,0)$	0	0	0	0	1	1	0	0
$\{1,2\},(1,1)$	0	0	0	0	0	0	1	1
$\{2,3\},(0,0)$	1	0	0	0	1	0	0	0
$\{2,3\},(0,1)$	0	1	0	0	0	1	0	0
$\{2,3\},(1,0)$	0	0	1	0	0	0	1	0
$\{2,3\},(1,1)$	0	0	0	1	0	0	0	1
$\{1,3\},(0,0)$	1	0	1	0	0	0	0	0
$\{1,3\},(0,1)$	0	1	0	1	0	0	0	0
$\{1,3\},(1,0)$	0	0	0	0	1	0	1	0
$\{1,3\},(1,1)$	0	0	0	0	0	1	0	1

.

Kahle has shown in [22], Theorem 14, that the set of support sets of distributions in the closure $\overline{\mathcal{E}^2}$ includes all subsets of size at most three. Thus the equi-distribution on $Y := \{(0, 0, 1), (0, 1, 0), (1, 0, 0)\}$ lies in $\overline{\mathcal{E}^2}$. On the other hand Y is not feasible with respect to $\binom{[3]}{2}$. So the equi-distribution on Y does not factor according to $\binom{[3]}{2}$, it is not a product of the form (5). We remark that Develin and Sullivant [15] have shown that the non-negative toric variety cutting $\overline{\mathcal{E}^2}$ out of the probability simplex $\Delta(X)$ is defined by the polynomial equation

$$p(0,0,0)p(0,1,1)p(1,0,1)p(1,1,0) = p(0,0,1)p(0,1,0)p(1,0,0)p(1,1,1).$$

Example 2.1 generalizes to higher-order Gibbs families.

Example 2.2. Let $k, N \in \mathbb{N}$ and $N > k \geq 2$. Then there exist distributions in the closure $\overline{\mathcal{E}^k}$ of $\mathcal{E}^k$, the set of Gibbs states of maximal order k, that do not factor according to $\binom{[N]}{k}$. We consider for $k \in \mathbb{N}$ a number of $N = k+1$ binary units. The subset $Y \subset X = \{0,1\}^N$ of size N, defined by

$$Y := \{(x_1, \ldots, x_N) \mid x_i = 0 \text{ for all but one } i \in [N]\}$$

is not feasible with respect to $\binom{[N]}{k}$. So the equi-distribution on Y has no product form (5). On the other hand, the support sets of distributions in $\overline{\mathcal{E}^k}$ include all subsets of size $2^k - 1$ by Theorem 14 in [22]. Since $2^k - 1 \geq k + 1$ holds for $k \geq 2$ by the Bernoulli inequality, the equi-distribution on Y lies in $\overline{\mathcal{E}^k}$.

The cardinality of non-feasible sets in Example 2.2 is minimal. In particular, the uniform distribution on every non-empty subset of X of cardinality at most k factors according to $\binom{[N]}{k}$.

Lemma 2.3. *Let $l, k, N \in \mathbb{N}$ and $1 \leq l \leq k \leq N$. Then every subset of X of cardinality l is feasible with respect to $\binom{[N]}{k}$.*

Proof: For $x \in X$ we denote the support $\operatorname{supp}(a_x)$ of the x-th column of A^k by $\operatorname{supp}^k(x)$ where A^k is the matrix of $\binom{[N]}{k}$ defined in (6). Let $Y \subset X$ be any subset of cardinality l and let $z \in X \setminus Y$. If $l \geq 2$ then we have

$$\operatorname{supp}^k(z) \subset \bigcup_{y \in Y} \operatorname{supp}^k(y) \quad \Longrightarrow \quad \forall x \in Y : \operatorname{supp}^{k-1}(z) \subset \bigcup_{y \in Y \setminus \{x\}} \operatorname{supp}^{k-1}(y). \tag{7}$$

We give a proof by contradiction. The conclusion of (7) says by definition that for all $x \in Y$ and all $A \subset [N]$ of size $k-1$ there exists $y \in Y \setminus \{x\}$ such that $z_A = y_A$. The negation asserts the existence of $x \in Y$ and $A \subset [N]$ of size $k-1$ such that for all $y \in Y \setminus \{x\}$ we have $z_A \neq y_A$. Hence, for all subsets $B \subset [N]$, $B \supset A$ of size k and for all $y \in Y \setminus \{x\}$ we have $z_B \neq y_B$. The premise of (7) then shows $z_B = x_B$. Since one point of B, the one not in A, is free to move within $[N]$, we get $z = x$ and the contradiction $z \in Y$ follows.

Now, if a subset $Y \subset X$ of cardinality l is not feasible with respect to $\binom{[N]}{k}$ then there exists $z \in X \setminus Y$ such that the premise of (7) is true. Applying (7) $l-1$ times shows for all $x \in Y$ that $\operatorname{supp}^{k-l+1}(z) = \operatorname{supp}^{k-l+1}(x)$ holds. Since $k - l + 1 \geq 1$ holds, this proves $z = x$ and contradicts $z \notin Y$. □

3. Fundamentals

We now define Gibbs families of quantum states. Since the infimum divergence from a Gibbs family is not achieved in general, we have to introduce a topological closure. This closure turns out to be equal to all maximum-entropy states under expected value constraints of the Hamiltonians. Moreover, the infimum divergence from a Gibbs family is achieved on this closure so that the divergence from a Gibbs family characterizes maximum-entropy states. The well-known decomposition of divergence in the presence of an increasing sequence of Gibbs families extends to the closure. Finally, we show what can go wrong: The divergence from a Gibbs family is discontinuous in the quantum case because of closure discrepancies in different

topologies. Unfortunately also, the divergence does not characterize the factoring property of Gibbs states of maximal order k, already classically.

We describe an m-level quantum system, $m \in \mathbb{N}$, in terms of a suitable complex C*-algebra. More precisely, we denote the C*-algebra of complex $m \times m$-matrices by $\mathcal{M}_m$ and call $\mathcal{A} \subset \mathcal{M}_m$ a *subalgebra* of $\mathcal{M}_m$, if $\mathcal{A}$ is a complex C*-algebra including the $m \times m$-identity matrix $\mathbb{1}_m$, also denoted by $\mathbb{1}_{\mathcal{A}}$. A subalgebra $\mathcal{A}$ is a unitary space with the Hilbert-Schmidt inner product $\langle a, b\rangle := \mathrm{tr}(a^*b)$, $a, b \in \mathcal{A}$, where tr denotes the standard trace. We denote the Euclidean space of self-adjoint matrices, the *observables*, by $\mathcal{A}|_{\mathrm{sa}} := \{a \in \mathcal{A} \mid a^* = a\}$, endowed with the restricted inner product $\langle a, b\rangle = \mathrm{tr}(ab)$. The state of a quantum system is described by a *density matrix* in $\mathcal{A}$, that is a positive semi-definite matrix of trace one. The compact convex set of density matrices, denoted by $\mathcal{S}_{\mathcal{A}}$, is called *state space* of the m-level system. In the classical case, $\mathcal{A} \cong \mathbb{C}^m$ is the algebra of diagonal matrices and $\mathcal{S}_{\mathcal{A}} \cong \Delta([m])$ is the probability simplex of dimension $m-1$. The state space $\mathcal{S}_{\mathcal{M}_m}$ has real dimension $m^2 - 1$.

If a self-adjoint matrix $h_0 \in \mathcal{A}|_{\mathrm{sa}}$ and a real subspace $H \subset \mathcal{A}|_{\mathrm{sa}}$ are chosen, we use the map $\mathcal{A}|_{\mathrm{sa}} \to \mathcal{S}_{\mathcal{A}}$, $R(a) = e^a/\mathrm{tr}\, e^a$, and define a *Gibbs family* by $R(h_0 + H)$. In statistical physics, the elements of H are called *Hamiltonians* or *energies*. The divergence from a Gibbs family $\mathcal{E}$ will be denoted

$$\mathrm{d}_{\mathcal{E}}(\rho) := \inf\{D(\rho\|\sigma) \mid \sigma \in \mathcal{E}\}, \quad \rho \in \mathcal{S}_{\mathcal{A}}. \tag{8}$$

The easiest example where the infimum is not achieved is the divergence of a state of reduced support from the Gibbs family of all invertible density matrices $R(\mathcal{A}|_{\mathrm{sa}})$.

The divergence from a Gibbs family was described in terms of topology and geometry by Weis [42]. *Information neighborhoods* $\{\sigma \in \mathcal{S}_{\mathcal{A}} \mid D(\rho\|\sigma) < \epsilon\}$, $\rho \in \mathcal{S}_{\mathcal{A}}$, for $\epsilon \in (0, \infty]$, are the basis of a topology on $\mathcal{S}_{\mathcal{A}}$, called *rI-topology* (the acronym *rI* stands for *reverse information* and *reverse* refers to the argument order of the divergence). The convergence of a sequence $(\rho_i)_{i\in\mathbb{N}} \subset \mathcal{S}_{\mathcal{A}}$ to $\rho \in \mathcal{S}_{\mathcal{A}}$ in the rI-topology is equivalent to $\lim_{i\to\infty} D(\rho\|\rho_i) = 0$, see [42]. We then say the sequence *rI-converges*. The closure of a subset $X \subset \mathcal{S}_{\mathcal{A}}$ in the rI-topology is equal to the *rI-closure*

$$\mathrm{cl}^{\mathrm{rI}}(X) := \{\rho \in \mathcal{S}_{\mathcal{A}} \mid \inf_{\sigma\in X} D(\rho\|\sigma) = 0\}$$

and consists of all limit points of rI-converging sequences in X.

Let us recall that analogous information neighborhoods in infinite-dimensional C*-algebras do not define a topology because of negative results shown by Csiszár [12] about infinite σ-algebras. Harremoës [18] has shown for infinite σ-algebras that the convergence with respect to the Kullback-Leibler divergence is more useful than information neighborhoods. Shirokov has confirmed the utility of information convergence for density matrices on a separable Hilbert space, see [37], Proposition 2.

In our setting of a C*-algebra $\mathcal{A}$ of finite-dimensional matrices, geometrical properties of the rI-closure $\mathrm{cl}^{\mathrm{rI}}(\mathcal{E})$ of a Gibbs family $\mathcal{E} = R(h_0 + H)$ are studied in Section 3 and Section 6 in [42]. For every state $\rho \in \mathcal{S}_{\mathcal{A}}$ exists a unique state in $\mathrm{cl}^{\mathrm{rI}}(\mathcal{E})$, denoted $\pi_{\mathcal{E}}(\rho)$, such that $\langle h, \rho\rangle = \langle h, \pi_{\mathcal{E}}(\rho)\rangle$ holds for all $h \in H$. Recall that the real number $\langle a, \rho\rangle$ is interpreted in quantum mechanics as the expected value of the observable a if the system is in the state ρ. The *projection theorem* says for every $\rho \in \mathcal{S}_{\mathcal{A}}$ that we have

$$\mathrm{d}_{\mathcal{E}}(\rho) = D(\rho\|\pi_{\mathcal{E}}(\rho)) = \min\{D(\rho\|\sigma) \mid \sigma \in \mathrm{cl}^{\mathrm{rI}}(\mathcal{E})\}. \tag{9}$$

The *Pythagorean theorem* says for every $\rho \in \mathcal{S}_{\mathcal{A}}$ and for every $\sigma \in \mathrm{cl}^{\mathrm{rI}}(\mathcal{E})$ that

$$D(\rho\|\sigma) = D(\rho\|\pi_{\mathcal{E}}(\rho)) + D(\pi_{\mathcal{E}}(\rho)\|\sigma) \tag{10}$$

holds. The theorems (9) and (10) are topological extensions of results in information geometry, see for example Petz [33] or Amari and Nagaoka [2] and non-commutative extensions of results in probability theory, see for example Barndorff-Nielsen [8] or Csiszár and Matúš [13].

The linear case $h_0 = 0$ of a Gibbs family $\mathcal{E} = R(H)$ solves a maximum-entropy problem under expected value constraints because we can take $\sigma = \mathbb{1}_{\mathcal{A}}/\mathrm{tr}(\mathbb{1}_{\mathcal{A}})$ in (10). This shows for $\rho \in \mathcal{S}$

$$\mathrm{argmax}\{H(\tau) \mid \tau \in \mathcal{S}_{\mathcal{A}}, \forall h \in H : \langle h, \tau\rangle = \langle h, \rho\rangle\} = \pi_{\mathcal{E}}(\rho). \tag{11}$$

See Section 3.4 in [42] for the details of this easy calculation.

Let us turn to the decomposition of divergences. Given an increasing sequence of real subspaces $H_1 \subset H_2 \subset \cdots \subset H_k \subset \mathcal{A}|_{\mathrm{sa}}$, for $k \in \mathbb{N}$, we consider Gibbs families $\mathcal{E}_i := R(H_i)$, $i = 1, \ldots, k$. The definition of the projection $\pi_{\mathcal{E}}$ shows $\pi_{\mathcal{E}_i} \circ \pi_{\mathcal{E}_j} = \pi_{\mathcal{E}_i}$ for all $1 \leq i \leq j \leq k$ and (10) shows for all $\rho \in \mathcal{S}_{\mathcal{A}}$

$$D(\rho\|\pi_{\mathcal{E}_i}(\rho)) = D(\rho\|\pi_{\mathcal{E}_j}(\rho)) + D(\pi_{\mathcal{E}_j}(\rho)\|\pi_{\mathcal{E}_i}(\rho)).$$

By induction we have for $1 \leq i \leq k$ and all $\rho \in \mathcal{S}_{\mathcal{A}}$

$$D(\rho\|\pi_{\mathcal{E}_i}(\rho)) = D(\rho\|\pi_{\mathcal{E}_k}(\rho)) + \sum_{j=i}^{k-1} D(\pi_{\mathcal{E}_{j+1}}(\rho)\|\pi_{\mathcal{E}_j}(\rho)),$$

which by (9) is the same as

$$\mathrm{d}_{\mathcal{E}_i}(\rho) = \mathrm{d}_{\mathcal{E}_k}(\rho) + \sum_{j=i}^{k-1} \mathrm{d}_{\mathcal{E}_j}(\pi_{\mathcal{E}_{j+1}}(\rho)). \tag{12}$$

Let us emphasize that the rI-topology is strictly finer (has more open sets) than the norm topology on the state space $\mathcal{S}_{\mathcal{A}}$ of a non-commutative algebra $\mathcal{A}$, see Corollary 5.19 in [42]. This finds expression in the rI-closure of some Gibbs families such as the example defined by two block-diagonal matrices $h_1 := \sigma_1 \oplus 0$, $h_2 := \sigma_1 \oplus 1$, for Pauli matrices $\sigma_1 := \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ and $\sigma_2 := \left(\begin{smallmatrix} 0 & -i \\ i & 0 \end{smallmatrix}\right)$. Weis and Knauf [43] have shown that the Gibbs family $\{R(\lambda h_1 + \mu h_2) \mid \lambda, \mu \in \mathbb{R}\}$ has an rI-closure which is strictly included in the norm closure. The divergence from this Gibbs family is discontinuous, see Section IV.B in [43] or Section 6.6 in [42].

Unfortunately, unlike the geometric properties (9) and (10), the maximum-entropy property (11) and the decomposition (12), the issue of factoring (5) is not characterized by the divergence. Already classically, the divergence $\mathcal{I}^k = \mathrm{d}_{\mathcal{E}^k}$ from $\mathcal{E}^k$, the Gibbs states of maximal order k, does not characterize factoring according to $\binom{[N]}{k}$ for $k \geq 2$. In fact, in the simplex of probability distributions $\Delta([m])$ which is the state space of the commutative algebra $\mathcal{A} \cong \mathbb{C}^m$ of diagonal matrices the rI-topology equals the norm topology by Corollary 5.19 in [42]. Hence, the projection theorem (9) shows for all distributions p in the norm closure of $\mathcal{E}^k$ that

$$\mathcal{I}^k(p) = \mathrm{d}_{\mathcal{E}^k}(p) = 0$$

holds. For $k = 2$ the equi-distribution p on $\{(0,0,1), (0,1,0), (1,0,0)\}$, unlike all distributions in $\mathcal{E}^2$, does not factor into a product $p(x,y,z) = \Psi_{\{1,2\}}(x,y) \cdot \Psi_{\{1,3\}}(x,z) \cdot \Psi_{\{2,3\}}(y,z)$, $x, y, z \in \{0,1\}$, by Example 2.1. Nevertheless p belongs to the closure

of $\mathcal{E}^2$. For $k > 2$ we can argue analogously using Example 2.2. While classically this inconsistency between factoring and zero-divergence is a boundary phenomenon, an analogous factoring property seems to be completely missing in the non-commutative setting except for $k = 1$.

4. Hierarchical models of quantum states

We define hierarchical models of quantum states, we calculate their dimensions and we discuss adapted bases of observables. The notion of k-body potentials is classical in statistical mechanics, see e.g. Ruelle [36], Chapter 2.4. Somewhat more general we consider Hamiltonians of order k. For example, Hamiltonians h of maximal order two define quantum Boltzmann machines $e^h/\mathrm{tr}(e^h)$, see Yapage and Nagaoka [44]. Similar concepts appear in theoretical biology and other disciplines, and have been abstractly studied under the name of *hierarchical model*, see Lauritzen [28], Chapter 4.3 and Appendix B.2.

The set $[N] = \{1, \dots, N\}$ for $N \in \mathbb{N}$ enumerates the *units* of a composite system. We now consider subalgebras $\mathcal{A}_k$, $k \in [N]$, of algebras of complex square matrices of arbitrary finite sizes, and the C^*-algebra

$$\mathcal{A} := \bigotimes_{k\in[N]} \mathcal{A}_k.$$

To a non-empty subset $v \subset [N]$ we associate the *factor space* $\mathcal{F}_v := \bigotimes_{k\in v} \mathcal{A}_k$, embedded into $\mathcal{A}$ by tensoring with $\bigotimes_{k\in[N]\setminus v} \mathbb{1}_{\mathcal{A}_k}$, and we set $\mathcal{F}_\emptyset := \mathrm{span}_{\mathbb{C}}(\mathbb{1}_{\mathcal{A}})$. So $\dim_{\mathbb{C}}(\mathcal{F}_v) = \prod_{k\in v} \dim_{\mathbb{C}}(\mathcal{A}_k)$, and $\mathcal{F}_w \subset \mathcal{F}_v$ for $w \subset v$. The *pure factor space* $\tilde{\mathcal{F}}_v \subset \mathcal{F}_v$ is then defined to be the maximal subspace orthogonal (w.r.t. Hilbert-Schmidt scalar product) to all $\mathcal{F}_w$ with $w \subsetneq v$. So $\mathcal{F}_v = \bigoplus_{w\subset v} \tilde{\mathcal{F}}_w$, and by Möbius inversion applied to the dimensions of the subspaces, see for example Appendix A.3 in [28],

$$\dim_{\mathbb{C}}(\tilde{\mathcal{F}}_v) = \prod_{k\in v} \big(\dim_{\mathbb{C}}(\mathcal{A}_k) - 1\big). \tag{13}$$

An orthonormal basis of $\mathcal{A}$, compatible with the decomposition $\mathcal{A} = \bigoplus_{v\in 2^{[N]}} \tilde{\mathcal{F}}_v$ is any family of orthonormal bases $B^{(k)}$ of $\mathcal{A}_k$ indexed by the units $k \in [N]$ such that $\frac{\mathbb{1}_{\mathcal{A}_k}}{\sqrt{\mathrm{tr}\mathbb{1}_{\mathcal{A}_k}}} \in B^{(k)}$. Then

$$\left\{ \bigotimes_{k=1}^{N} b_k \mid b_m \in B^{(m)}, m \in [N] \right\}$$

is an orthonormal basis of $\mathcal{A}$ and for $v \subset [N]$ we have

$$\tilde{\mathcal{F}}_v = \mathrm{span} \left\{ \bigotimes_{k=1}^{N} b_k \mid b_m = \frac{\mathbb{1}_{\mathcal{A}_m}}{\sqrt{\mathrm{tr}\mathbb{1}_{\mathcal{A}_m}}} \text{ iff } m \notin v, b_m \in B^{(m)}, m \in [N] \right\}.$$

Sometimes a concrete basis is needed. For a full matrix algebra $\mathcal{M}_n$ we can use for $k, l = 0, \dots, n-1$ the matrices given (for $r, s = 1, \dots, n$) by

$$\left(E^{(n)}_{k,l}\right)_{r,s} := \tfrac{1}{\sqrt{n}} \Big(\exp\left(\pi i (r+s) \tfrac{k}{n}\right) \, \delta_{r-s+l} + \exp\left(\pi i (r+s-n) \tfrac{k}{n}\right) \delta_{r-s+l-n} \Big).$$

Lemma 4.1. $\left\{E^{(n)}_{k,l} \mid k,l \in \{0,\dots,n-1\}\right\} \subset \mathcal{M}_n$ *is an orthonormal basis of* $\mathcal{M}_n$. *The adjoints are* $E^{(n)*}_{k,0} = E^{(n)}_{n-k,0}$, $E^{(n)*}_{0,l} = E^{(n)}_{0,n-l}$ *and* $E^{(n)*}_{k,l} = (-1)^{n+k+l} E^{(n)}_{n-k,n-l}$ *for* $k,l = 1,\dots,n-1$.

Proof: For $k,l,k',l' \in \{0,\dots,n-1\}$

$$\begin{aligned}
\left\langle E^{(n)}_{k,l}, E^{(n)}_{k',l'} \right\rangle &= \sum_{r,s=1}^{n} \left(E^{(n)}_{k,l}\right)_{r,s} \overline{\left(E^{(n)}_{k',l'}\right)_{r,s}} \\
&= \frac{1}{n} \sum_{r,s=1}^{n} \Big[\exp\left(\pi\, i\, (r+s)(k-k')/n\right) \delta_{r-s+l}\, \delta_{r-s+l'} \;+ \\
&\qquad \exp\left(\pi\, i\, (r+s-n)(k-k')/n\right) \delta_{r-s+l-n}\, \delta_{r-s+l'-n} \Big] \\
&= \frac{1}{n}\, \delta_{l,l'} \sum_{r=1}^{n} \exp\left(\pi\, i\, (2r+l)(k-k')/n\right) = \delta_{l,l'}\, \delta_{k,k'}.
\end{aligned}$$

As the set has size n^2, this shows the claim. The following adjoints appear. One has $E^{(n)}_{0,0} = \frac{1}{\sqrt{n}} \mathbb{1}_n$. For $k = 1,\dots,n-1$ and coefficients $r,s = 1,\dots,n$

$$\begin{aligned}
\left(E^{(n)}_{k,0}\right)^*_{r,s} &= \frac{1}{\sqrt{n}} \overline{\exp\left(\pi i (r+s)k/n\right)} \delta_{r-s} = \frac{1}{\sqrt{n}} \exp\left(-\pi i (r+s)k/n\right) \delta_{r-s} \\
&= \frac{1}{\sqrt{n}} \exp\left(\pi i (r+s)(n-k)/n\right) \delta_{r-s} = \left(E^{(n)}_{n-k,0}\right)_{r,s}
\end{aligned}$$

holds and for $l = 1,\dots,n-1$ it is immediate that $E^{(n)*}_{0,l} = E^{(n)}_{0,n-l}$. For $k,l = 1,\dots,n-1$ and coefficients $r,s = 1,\dots,n$ one has

$$\begin{aligned}
\left(E^{(n)}_{k,l}\right)^*_{r,s} &= \frac{1}{\sqrt{n}} \Big(\exp\left(-\pi i (r+s-n)k/n\right) \delta_{r-s+n-l} \;+ \\
&\qquad \exp\left(-\pi i (r+s)k/n\right) \delta_{r-s-l} \Big) \\
&= \frac{1}{\sqrt{n}} \Big((-1)^{k+r+s} \exp\left(\pi i (r+s)(n-k)/n\right) \delta_{r-s+n-l} \;+ \\
&\qquad (-1)^{n+k+r+s} \exp\left(\pi i (r+s-n)(n-k)/n\right) \delta_{r-s-l} \Big) \\
&= (-1)^{n+k+l} \left(E^{(n)}_{n-k,n-l}\right)_{r,s}.
\end{aligned}$$

□

One way to compute a self-adjoint basis out of the basis $\{E^{(n)}_{k,l}\}_{k,l=0}^{n-1}$ of $\mathcal{M}_n$, $n \in \mathbb{N}$, in Lemma 4.1, is to use their symmetry under hermitian conjugation. Orbits have length one or two. Thus the transformation of basis matrices E to pairs of matrices $E + E^*$ and $i(E - E^*)$ produces exactly n^2 pairwise orthogonal non-zero self-adjoint matrices. This symmetrization is different compared to the basis (3.2) by Petz [33], where only real hermitian matrices appear which are either diagonal or which have only two non-zero entries. In contrast

$$E^{(3)}_{0,1} + (E^{(3)}_{0,1})^* = \frac{1}{\sqrt{3}} \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$

Coming back to the subject of hierarchical models, let $U \subset 2^{[N]}$ be a class of subsets of $[N]$. Differring from common terminology, we will call U a *hypergraph* on $[N]$ if

$$v \in U,\ w \subset v \ \ \Rightarrow \ \ w \in U, \qquad \text{and} \qquad \bigcup_{v \in U} v = [N].$$

We consider a hypergraph U on $[N]$ and define $\tilde{\mathcal{F}}_U := \bigoplus_{v \in U} \tilde{\mathcal{F}}_v$. A *hierarchical model* $\mathcal{E}_U$ of a hypergraph U on $[N]$ is defined as the Gibbs family

$$\mathcal{E}_U := R(\tilde{\mathcal{F}}_U \cap \mathcal{A}|_{\mathrm{sa}}). \tag{14}$$

Of particular interest are the hypergraphs $U^k = \bigcup_{\ell=0}^{k} \binom{[N]}{\ell}$. Similarly to the classical case in Section 2 in the paragraph of (4), the *order* of a self-adjoint matrix $a \in \mathcal{A}|_{\mathrm{sa}}$ is the minimal number $k \in \mathbb{N}$ such that $a \in \bigoplus_{i=0}^{k} \tilde{\mathcal{F}}_{U^i}$. In the context of Gibbs states $R(a) = e^a/\mathrm{tr}(e^a)$ we call a self-adjoint matrix $a \in \mathcal{A}|_{\mathrm{sa}}$ *Hamiltonian* and we define the *order* of $R(a)$ equal to the order of a. In quantum complexity theory [26, 14] a Hamiltonian of order at most $k \in \mathbb{N}$ is known as *k-local Hamiltonian*. The set $\mathcal{E}^k$ of *Gibbs states of maximal order k* is the hierarchical model of the hypergraph U^k. For example, the *independence model* $\mathcal{E}^1$ is the hierarchical model of the hypergraph $\{\emptyset, \{1\}, \ldots, \{N\}\}$.

This formal definition of $\mathcal{E}^k$ was anticipated in the introduction where the functionals $\mathcal{I}^k(\rho) = \inf\{D(\rho\|\sigma) \mid \sigma \in \mathcal{E}^k\}$, $\rho \in \mathcal{S}_{\mathcal{A}}$, are defined in (1). As mentioned in the beginning of this section, the order of a quantum state is basic in statistical mechanics. The functional $\mathcal{I}^k$ quantifies the higher-order correlations in a state which do not arise from Hamiltonians of maximal order k. Strictly speaking, see Section 3, one should be more careful because of states of reduced support having zero divergence $\mathcal{I}^k$. Similarly, a measure

$$\mathcal{I}^2 - \mathcal{I}^3, \tag{15}$$

defined by Kahle et al. [23], quantifies those correlations between triples of units that are not captured by correlations between pairs of units. This, too, is not a strictly correct interpretation but it is reasonable for all practical purposes.

We now compute dimensions of hierarchical models. The relative interior of a subset of $\mathcal{A}|_{\mathrm{sa}}$ is the interior of the subset in its affine hull.

Proposition 4.2. *The Gibbs family $\mathcal{E}_U \subset \mathcal{S}_{\mathcal{A}}$ is a manifold of dimension*

$$\dim_{\mathbb{R}}(\mathcal{E}_U) = \sum_{\substack{v \in U \\ v \neq \emptyset}} \prod_{i \in v} \big(\dim_{\mathbb{C}}(\mathcal{A}_i) - 1\big).$$

Proof: By the definition of hypergraphs and (13) we have for all $v \in 2^{[N]}$

$$\dim_{\mathbb{C}}\big(\tilde{\mathcal{F}}_v\big) = \prod_{i \in v} \big(\dim_{\mathbb{C}}(\mathcal{A}_i) - 1\big).$$

Notice that every complex *-invariant subspace of $\mathcal{A}$ is a direct sum of two copies of the real subspace of self-adjoint elements. Therefore

$$\dim_{\mathbb{R}}(\tilde{\mathcal{F}}_v \cap \mathcal{A}|_{\mathrm{sa}}) = \dim_{\mathbb{C}}(\tilde{\mathcal{F}}_v).$$

By definition, the hypergraph U contains $\emptyset$ and $\tilde{\mathcal{F}}_U = \tilde{\mathcal{F}}_\emptyset \oplus V$ is the direct sum of $\tilde{\mathcal{F}}_\emptyset = \mathrm{span}_{\mathbb{C}}(\mathbb{1}_{\mathcal{A}})$ and of its orthogonal complement, denoted V. Clearly $R(\tilde{\mathcal{F}}_U \cap \mathcal{A}|_{\mathrm{sa}}) = R(V \cap \mathcal{A}|_{\mathrm{sa}})$ holds. If $W \subset \mathcal{A}|_{\mathrm{sa}}$ is a codimension one subspace not containing

the identity $\mathbb{1}_{\mathcal{A}}$, then $R|_W$ is a diffeomorphism to the relative interior of $\mathcal{S}_{\mathcal{A}}$, see Proposition 6.1.2 in [42]. Then

$$\dim_{\mathbb{R}}(\mathcal{E}_U) = \dim_{\mathbb{R}}(V) = \sum_{\substack{v\in U\\ v\neq\emptyset}} \dim_{\mathbb{R}}(\tilde{\mathcal{F}}_v \cap \mathcal{A}|_{\mathrm{sa}})$$

completes the proof. □

5. The multi-information

We continue with the composite system from last section, having the tensor product algebra $\mathcal{A} := \bigotimes_{i=1}^N \mathcal{A}_i$ of algebras $\mathcal{A}_i$ associated to the individual units $i \in [N]$. It is an easy calculation to show that the *multi-information*

$$\mathcal{I}(\rho) := \sum_{i\in[N]} H(\rho^i) - H(\rho), \quad \rho \in \mathcal{S}_{\mathcal{A}} \tag{16}$$

equals the divergence $D(\rho\|\rho^1 \otimes \cdots \otimes \rho^N)$ of ρ from the product of marginal states ρ^i, defined implicitly by $\mathrm{tr}(X\rho^i) = \mathrm{tr}(X\rho)$, $X \in \mathcal{A}_i|_{\mathrm{sa}}$, where X is embedded into $\mathcal{A}$ by tensoring with $\bigotimes_{k\in[N]\setminus\{i\}} \mathbb{1}_{\mathcal{A}_k}$, $i \in [N]$. This equation shows that $\mathcal{I}(\rho)$ is zero only if ρ is a completely uncorrelated product state.

Modi et al. [29] have proved the minimality results of $D(\rho\|\rho^1\otimes\cdots\otimes\rho^N) \leq D(\rho\|\sigma)$ valid for all product states $\sigma \in \mathcal{P}$,

$$\mathcal{P} := \{\rho_1 \otimes \cdots \otimes \rho_N \mid \rho_i \in \mathcal{S}_{\mathcal{A}_i}, i \in [N]\}, \tag{17}$$

with equality if and only if σ is the product of marginals $\rho^1 \otimes \cdots \otimes \rho^N$. This minimality is almost but not exactly the equality $\mathcal{I}^1 = \mathcal{I}$ between the divergence $\mathcal{I}^1 = \mathrm{d}_{\mathcal{E}^1}$ from the independence model $\mathcal{E}^1$ and multi-information $\mathcal{I}$. A proof of $\mathcal{I}^1 = \mathcal{I}$ was given in [40] using a construction of the rI-closure of $\mathcal{E}^1$ in terms of a union of independence models of reduced support. Here we give a shorter proof using the rI-topology defined in Section 3.

Theorem 5.1. *The set of product states $\mathcal{P}$ is the closure of the independence model $\mathcal{E}^1$ both in the rI- and norm topology, that is $\mathcal{P} = \mathrm{cl}^{\mathrm{rI}}(\mathcal{E}^1) = \overline{\mathcal{E}^1}$. The divergence $\mathcal{I}^1 = \mathrm{d}_{\mathcal{E}^1}$ from the independence model $\mathcal{E}^1$ is the multi-information $\mathcal{I}$.*

Proof: We prove $\mathcal{P} \subset \mathrm{cl}^{\mathrm{rI}}(\mathcal{E}^1)$. Let $\rho = \rho_1 \otimes \cdots \otimes \rho_N$ be a product state. It is shown in Theorem 5.18.5 in [42] that each individual factor ρ_i lies in the rI-closure of the relative interior of the state space $\mathcal{S}_{\mathcal{A}_i}$, which is the set of all invertible density matrices in $\mathcal{S}_{\mathcal{A}_i}$. So there exist sequences $(\rho_i^{(n)})_{n\in\mathbb{N}} \subset \mathcal{S}_{\mathcal{A}_i}$ of invertible states such that $\lim_{n\to\infty} D(\rho_i\|\rho_i^{(n)}) = 0$, $i \in [N]$. It follows

$$D\big(\rho\|\rho_1^{(n)} \otimes \cdots \otimes \rho_N^{(n)}\big) = D\big(\rho_1\|\rho_1^{(n)}\big) + \cdots + D\big(\rho_N\|\rho_N^{(n)}\big) \overset{n\to\infty}{\longrightarrow} 0.$$

Since $\rho_1^{(n)} \otimes \cdots \otimes \rho_N^{(n)} \in \mathcal{E}^1$ for all $n \in \mathbb{N}$ this proves $\rho \in \mathrm{cl}^{\mathrm{rI}}(\mathcal{E}^1)$. The inclusion $\mathrm{cl}^{\mathrm{rI}}(\mathcal{E}^1) \subset \overline{\mathcal{E}^1}$ is clear because the rI-topology is finer than the norm topology. The inclusion $\overline{\mathcal{E}^1} \subset \mathcal{P}$ follows because $\mathcal{E}^1 \subset \mathcal{P}$ and because $\mathcal{P}$ is norm closed (image of a compactum under a continuous map). This completes the proof of $\mathcal{P} = \mathrm{cl}^{\mathrm{rI}}(\mathcal{E}^1) = \overline{\mathcal{E}^1}$.

The divergence $\mathcal{I}^1(\rho) = \mathrm{d}_{\mathcal{E}^1}(\rho)$ of $\rho \in \mathcal{S}_\mathcal{A}$ from $\mathcal{E}^1$ is

$$\begin{aligned} \mathrm{d}_{\mathcal{E}^1}(\rho) &= \min\{D(\rho\|\sigma) \| \sigma \in \mathrm{cl}^{\mathrm{rI}}(\mathcal{E}^1)\} \\ &= \min\{D(\rho\|\sigma) \| \sigma \in \mathcal{P}\} \\ &= \mathcal{I}(\rho). \end{aligned}$$

In the first equality we have used the projection theorem (9). The second equality follows from the first part of this theorem. The third equality is the minimality shown by Modi et al. [29], see the discussion above. □

6. Local maximizers of the divergence

Two characterizations of a local maximizer of the divergence from a Gibbs family by Ay [5] generalize to the quantum setting. They have been developed by Weis and Knauf [43] and will now be used in their formulation in Section 6.5 in [42].

One of the characterizations is an upper bound on the support in terms of the dimension of the Gibbs family. To evaluate the bound we need the notion of *face* of the state space $\mathcal{S}_\mathcal{A}$, which denotes a convex subset $F \subset \mathcal{S}_\mathcal{A}$ such that every segment $\{(1-\lambda)x + \lambda y \mid 0 \leq \lambda \leq 1\}$ included in $\mathcal{S}_\mathcal{A}$ $(x, y \in \mathcal{S}_\mathcal{A})$ which meets F in an interior point $(0 < \lambda < 1)$ lies in F. Let $\rho \in \mathcal{S}_\mathcal{A}$ be a local maximizer of the divergence $\mathrm{d}_\mathcal{E}$ from the Gibbs family $\mathcal{E}$, defined in (8), and let F be the (unique) face of $\mathcal{S}_\mathcal{A}$ containing ρ in its relative interior. Then

$$\dim_\mathbb{R}(F) \leq \dim_\mathbb{R}(\mathcal{E})$$

holds by Proposition 6.17 in [42]. This bound follows from the strict convexity of the divergence on fibers of the projection $\pi_\mathcal{E} : \mathcal{S}_\mathcal{A} \to \mathrm{cl}^{\mathrm{rI}}(\mathcal{E})$ defined in Section 3. The dimension of F depends not only on ρ but also on the algebra $\mathcal{A}$.

Two extreme cases are discussed in Remark 6.18 in [42], the classical algebra $\mathcal{A} \cong \mathbb{C}^m$ of diagonal matrices for some $m \in \mathbb{N}$, where

$$\mathrm{rk}(\rho) \leq \dim_\mathbb{R}(\mathcal{E}) + 1 \tag{18}$$

and the full matrix algebra $\mathcal{A} := \mathcal{M}_m$, where

$$\mathrm{rk}(\rho) \leq \sqrt{\dim_\mathbb{R}(\mathcal{E}) + 1}. \tag{19}$$

To evaluate these bounds for a hierarchical model we assume unit sizes $n_i \in \mathbb{N}$, subalgebras $\mathcal{A}_i \subset \mathcal{M}_{n_i}$, $i \in [N]$, as well as a hypergraph $U \subset 2^{[N]}$ are given. Proposition 4.2 shows

$$\dim_\mathbb{R}(\mathcal{E}_U) = \sum_{\substack{v \in U \\ v \neq \emptyset}} \prod_{k \in v} \big(\dim_\mathbb{C}(\mathcal{A}_k) - 1\big).$$

In the classical case of diagonal matrices $\mathcal{A}_i \cong \mathbb{C}^{n_i}$, $i \in [N]$, the state space $\mathcal{S}_\mathcal{A}$ is the probability simplex $\Delta(X)$ on $X = [n_1] \times [n_2] \times \cdots \times [n_N]$. A probability distribution $p \in \Delta(X)$, which is a local maximizer of the divergence from the hierarchical model $\mathcal{E}_U$ satisfies by (18) the bound on its support size

$$|\mathrm{supp}(p)| \leq \sum_{\substack{v \in U \\ v \neq \emptyset}} \prod_{i \in v} (n_i - 1) + 1\,. \tag{20}$$

In the quantum case where all algebras $\mathcal{A}_i = \mathcal{M}_{n_i}$, $i \in [N]$, are full matrix algebras, a local maximizer $\rho \in \mathcal{S}_\mathcal{A}$ of the divergence from the hierarchical model $\mathcal{E}_U$ satisfies by (19) the rank bound

$$\mathrm{rk}(\rho) \leq \sqrt{\sum_{\substack{v \in U \\ v \neq \emptyset}} \prod_{i \in v} (n_i{}^2 - 1) + 1}\,. \tag{21}$$

The hierarchical model of U^k, see (4) and (14), is the set $\mathcal{E}^k$ of Gibbs states of maximal order k. Here we want to see the dependence on the number N of units for equally sized units $n_i = n \in \mathbb{N}$, $i \in [N]$. A local maximizer of the divergence $\mathcal{I}^k$ from $\mathcal{E}^k$ is in the classical case (20) constrained by

$$|\mathrm{supp}(p)| \leq \sum_{i=1}^{k} \binom{N}{i} (n-1)^i + 1 \tag{22}$$

and in the quantum case (21) by

$$\mathrm{rk}(\rho) \leq \sqrt{\sum_{i=1}^{k} \binom{N}{i} (n^2-1)^i + 1}. \tag{23}$$

For $k = 1$ (multi-information $\mathcal{I} = \mathcal{I}^1$) these rank bounds are $N(n-1)+1$ respectively $\sqrt{N(n^2-1)+1}$.

For curiosity we mention a second characterization of a local maximizer ρ of the divergence from a Gibbs family $\mathcal{E} = R(h_0 + H)$, defined in Section 3. Namely, ρ must have a special form. Let $\mathcal{A}$ be a subalgebra of $\mathcal{M}_m$, $m \in \mathbb{N}$. We recall that an orthogonal projection $p \in \mathcal{A}$ is a matrix such that $p = p^2 = p^*$ holds. As shown in Section 3.3 and Section 3.5 in [42], the state $\pi_\mathcal{E}(\rho) \in \mathrm{cl}^{\mathrm{rI}}(\mathcal{E})$ defined in Section 3 is of the form $qe^{qa_\rho q}/\mathrm{tr}(qe^{qa_\rho q})$ for some self-adjoint matrix $a_\rho \in h_0 + H$ and q is an orthogonal projection. Secondly, and surprisingly, the Corollary 6.19 in [42] shows that the local maximizer ρ of the divergence from $\mathcal{E}$ is itself of the form $\rho = pe^{pa_\rho p}/\mathrm{tr}pe^{pa_\rho p}$ for an orthogonal projection $p \in \mathcal{A}$. In the classical case this result was shown in [5] and means that a local maximizer p of the divergence from a Gibbs family $\mathcal{E}$ is equal to the conditional probability distribution $\pi_\mathcal{E}(p)(\,\cdot\,|\mathrm{supp}(p))$.

7. Global separable maximizers of the mutual information

Global maximizers of the multi-information of a composite system of $N \in \mathbb{N}$ units with product algebra $\mathcal{A} := \bigotimes_{i=1}^N \mathcal{A}_i$ were studied by Ay and Knauf [6]. For example, a classification was proved for global maximizers in the classical setting of diagonal matrices $\mathcal{A}_i \cong \mathbb{C}^{n_i}$, $n_i \in \mathbb{N}$, $i = 1, \ldots, N$. If the units are ordered by their size, such that $n_1 \leq \cdots \leq n_N$, then the classical bound of the multi-information (16) is

$$\mathcal{I}(p) \leq \sum_{i=1}^{N-1} \log(n_i), \quad p \in \mathcal{S}_\mathcal{A} \cong \Delta(n_1 \times \cdots \times n_N)$$

for probability distributions p. For example, two classical bits have $\log(2) = 1$ bit of maximal mutual information. The example of two maximally entangled qubits, for example the Bell state $\frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$, shows that quantum systems can break the classical bound. This is a reason why some of the basic ideas in [6] do not apply to the quantum setting of full matrix algebras, $\mathcal{A}_i = \mathcal{M}_{n_i}$, $i \in [N]$.

Here we show that some arguments from [6] are helpful in the maximization of multi-information on the separable states. A constrained form of this problem appears in the definition of classical correlations *à la* Groisman et al. [17]. By definition, a state in $\mathcal{A}$ is *separable* if it is a convex combination of product states $\rho_1 \otimes \cdots \otimes \rho_N$. A state which is not separable is *entangled* [31, 9]. We restrict the discussion to the simplest case of a bipartite system ($N = 2$) of two qubits $\mathcal{A}_1 = \mathcal{A}_2 = \mathcal{M}_2$ where the multi-information (16) is known as *mutual information*

$$\mathcal{I}(\rho) = H(\rho^1) + H(\rho^2) - H(\rho), \quad \rho \in \mathcal{S}_{\mathcal{A}}, \quad \mathcal{A} = \mathcal{A}_1 \otimes \mathcal{A}_2. \tag{24}$$

A state is *classically correlated á la* Modi et al. [29] if it can be diagonalized by local unitaries that is, matrices in the subgroup $U(2) \times U(2) \subset U(4)$.

Theorem 7.1. *For arbitrary separable two-qubit state ρ, its mutual information is bounded by $\mathcal{I}(\rho) \leq \log(2)$. The equality holds if and only if ρ is local unitary equivalent to $\frac{1}{2}(|0\rangle\langle 0| \otimes |0\rangle\langle 0| + |1\rangle\langle 1| \otimes |1\rangle\langle 1|)$.*

Proof: If ρ is separable, then $H(\rho^i) \leq H(\rho)$, $i = 1, 2$, holds, see [32]. So we have

$$\mathcal{I}(\rho) \leq \min\{H(\rho^1),\ H(\rho^2)\}. \tag{25}$$

For qubit states ρ^1 and ρ^2, the maximum of the von Neumann entropy is no more than $\log(2)$, which constrains the maximum of mutual information $\mathcal{I}(\rho)$ to $\log(2)$. So if $\mathcal{I}(\rho)$ reaches its maximum $\log(2)$, then $H(\rho^i)$, $i = 1, 2$, also reaches this maximum, which requires ρ^i to be the maximally mixed state $\frac{1}{2}\mathbb{1}_2$.

Two-qubit mixed states with maximally mixed reduced states are local unitary equivalent to *Bell-diagonal states*

$$\rho = \sum_{i=1}^{4} \lambda_i |\psi_i\rangle\langle\psi_i|, \quad \lambda_1, \lambda_2, \lambda_3, \lambda_4 \geq 0, \quad \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1 \tag{26}$$

with $|\psi_1\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$, $|\psi_2\rangle = \frac{1}{\sqrt{2}}(|00\rangle - |11\rangle)$, $|\psi_3\rangle = \frac{1}{\sqrt{2}}(|01\rangle + |10\rangle)$, $|\psi_4\rangle = \frac{1}{\sqrt{2}}(|01\rangle - |10\rangle)$, see [35]. Note that $-H(\rho)$ is a strictly convex function of quantum states, subsequently, the maximum of $\mathcal{I}(\rho)$ on the convex set of separable Bell-diagonal states is attained only on the extreme points of this convex set. A Bell-diagonal state is separable if and only if $\lambda_i \leq \frac{1}{2}$ for $i = 1, 2, 3, 4$, see [19, 27]. We find the extreme points of the set of separable Bell-diagonal states are

$$\frac{1}{2}(|\psi_i\rangle\langle\psi_i| + |\psi_j\rangle\langle\psi_j|), \quad i \neq j, \quad i, j = 1, 2, 3, 4. \tag{27}$$

One can verify further that the mutual information of all these extreme points is $\log(2)$. Therefore the separable two-qubit states with maximum mutual information are all local unitary equivalent to the quantum state in (27).

Now we take a closer look at these maximizers. We find they are all classically correlated, since

$$\begin{array}{rcl}
\frac{1}{2}(|\psi_1\rangle\langle\psi_1| + |\psi_2\rangle\langle\psi_2|) & = & \frac{1}{2}(|0\rangle\langle 0| \otimes |0\rangle\langle 0| + |1\rangle\langle 1| \otimes |1\rangle\langle 1|); \\
\frac{1}{2}(|\psi_1\rangle\langle\psi_1| + |\psi_3\rangle\langle\psi_3|) & = & \frac{1}{2}(|+\rangle\langle +| \otimes |+\rangle\langle +| + |-\rangle\langle -| \otimes |-\rangle\langle -|); \\
\frac{1}{2}(|\psi_1\rangle\langle\psi_1| + |\psi_4\rangle\langle\psi_4|) & = & \frac{1}{2}(|0'\rangle\langle 0'| \otimes |1'\rangle\langle 1'| + |1'\rangle\langle 1'| \otimes |0'\rangle\langle 0'|); \\
\frac{1}{2}(|\psi_2\rangle\langle\psi_2| + |\psi_3\rangle\langle\psi_3|) & = & \frac{1}{2}(|1'\rangle\langle 1'| \otimes |1'\rangle\langle 1'| + |0'\rangle\langle 0'| \otimes |0'\rangle\langle 0'|); \\
\frac{1}{2}(|\psi_2\rangle\langle\psi_2| + |\psi_4\rangle\langle\psi_4|) & = & \frac{1}{2}(|-\rangle\langle -| \otimes |+\rangle\langle +| + |+\rangle\langle +| \otimes |-\rangle\langle -|); \\
\frac{1}{2}(|\psi_3\rangle\langle\psi_3| + |\psi_4\rangle\langle\psi_4|) & = & \frac{1}{2}(|0\rangle\langle 0| \otimes |1\rangle\langle 1| + |1\rangle\langle 1| \otimes |0\rangle\langle 0|),
\end{array} \tag{28}$$

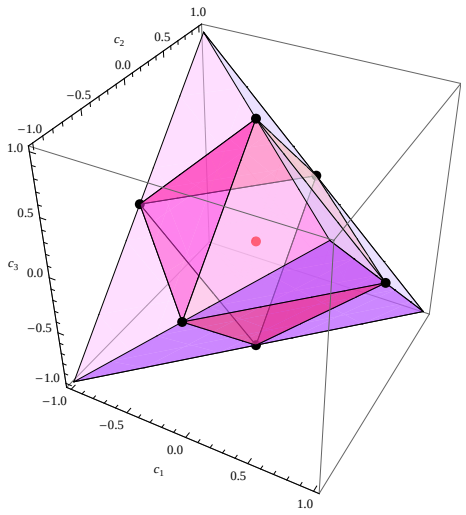

FIGURE 1. (Color online) Geometry of Bell-diagonal states.

with $|+\rangle = \frac{1}{\sqrt{2}}(|0\rangle+|1\rangle)$, $|-\rangle = \frac{1}{\sqrt{2}}(|0\rangle-|1\rangle)$, $|0'\rangle = \frac{1}{\sqrt{2}}(|0\rangle+i|1\rangle)$, $|1'\rangle = \frac{1}{\sqrt{2}}(|0\rangle-i|1\rangle)$. Here $\{|+\rangle, |-\rangle\}$ and $\{|0'\rangle, |1'\rangle\}$ are another two orthonormal bases of two dimensional Hilbert space. From equations (28) it is direct to get that all the maximizers are local unitary equivalent to $\frac{1}{2}(|0\rangle\langle 0| \otimes |0\rangle\langle 0| + |1\rangle\langle 1| \otimes |1\rangle\langle 1|)$. □

We finish with a geometric discussion of Theorem 7.1. Mutual information is the relative entropy of a quantum state from its closest product state, $\mathcal{I}(\rho) = \min_{\pi\in\mathcal{P}} D(\rho\|\pi)$, see (17). Hence, the mutual information $\mathcal{I}(\rho)$ can be regarded as the distance between a quantum state and the set of product states $\mathcal{P}$. In a two-qubit system, the maximum distance between an arbitrary separable quantum state and the set of product states $\mathcal{P}$ is $\log(2)$. Theorem 7.1 shows the farthest separable states from the set of product states $\mathcal{P}$ are all local unitary equivalent to $\frac{1}{2}(|0\rangle\langle 0| \otimes |0\rangle\langle 0| + |1\rangle\langle 1| \otimes |1\rangle\langle 1|)$. These states are classically correlated so they can not be used in the protocol of entanglement distribution *via* separable states in [25].

The Bell-diagonal states can be written as $\rho = \frac{1}{4}(\mathbb{1}_4 + \sum_{i=1}^{3} c_i\sigma_i \otimes \sigma_i)$ with σ_i three Pauli operators. So a Bell-diagonal state is specified by three real variables c_1, c_2, and c_3. One can show that a Bell-diagonal state is separable if and only if $|c_1| + |c_2| + |c_3| \leq 1$ holds. Geometrically, the set of Bell-diagonal states is a tetrahedron and the set of separable Bell-diagonal states is an octahedron, see [19, 27] and Fig. 1 for a drawing. The four vertices of the tetrahedron are Bell states $|\psi_i\rangle$ which are maximally entangled, $i = 1, 2, 3, 4$. The six black vertices of the octahedron are maximizers of the mutual information and they are classically correlated. The center red point $\frac{1}{4}\mathbb{1}_4$ is the only product state in this tetrahedron.

Acknowledgements. SW thanks Thomas Kahle for a helpful correspondence about factoring of probability distributions. SW was supported by the DFG projects "Geometry and Complexity in Information Theory" and "Quantum Statistics: Decision problems and entropic functionals on state spaces".

References

[1] Amari, S.-I. (2001). *Information geometry on hierarchy of probability distributions.* IEEE Transactions on Information Theory, 47(5), 1701–1711.

[2] Amari, S.-I., & Nagaoka, H. (2000). *Methods of information geometry.* Translations of Mathematical Monographs 191, American Mathematical Soc., Oxford University Press.

[3] Aoki, S., Hara, H., & Takemura, A. (2012). *Markov bases in algebraic statistics.* Springer Series in Statistics 199, Springer, New York.

[4] Ay, N. (2001). *Information geometry on complexity and stochastic interaction.* MIS-Preprint: 95/2001.

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

[6] Ay, N., & Knauf, A. (2006). *Maximizing multi-information.* Kybernetika, 42(5), 517–538.

[7] Ay, N., Olbrich, E., Bertschinger, N., & Jost, J. (2011). *A geometric approach to complexity.* Chaos, 21, 037103.

[8] Barndorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory.* Wiley, New York.

[9] Bengtsson, I., & Życzkowski, K. (2006). *Geometry of quantum states: an introduction to quantum entanglement.* Cambridge University Press.

[10] Benatti, F., Hudetz, T., & Knauf, A. (1998). *Quantum chaos and dynamical entropy.* Communications in Mathematical Physics, 198(3), 607–688.

[11] Cafaro, C., Giffin, A., Lupo, C., & Mancini, S. (2012). *Softening the complexity of entropic motion on curved statistical manifolds.* Open Systems & Information Dynamics, 19(1), 1250001.

[12] Csiszár, I. (1967). *On topological properties of f-divergences.* Studia Sci. Math. Hungar., 2, 329–339.

[13] Csiszár, I., & Matúš, F. (2003). *Information projections revisited.* IEEE Transactions on Information Theory, 49(6), 1474–1490.

[14] Cubitt, T., & Montanaro, A. (2014). *Complexity classification of local Hamiltonian problems.* arXiv:1311.3161 [quant-ph]

[15] Develin, M., & Sullivant, S. (2003). *Markov bases of binary graph models.* Annals of Combinatorics, 7(4), 441–466.

[16] Geiger, D., Meek, C., & Sturmfels, B. (2006). *On the toric algebra of graphical models.* The Annals of Statistics, 34(3), 1463–1492.

[17] Groisman, B., Popescu, S., & Winter, A. (2005). *Quantum, classical, and total amount of correlations in a quantum state.* Physical Review A, 72(3), 032317.

[18] Harremoës, P. (2007). *Information topologies with applications.* In Entropy, Search, Complexity (pp. 113–150). Springer Berlin Heidelberg.

[19] Horodecki, R., & Horodecki, M. (1996). *Information-theoretic aspects of inseparability of mixed states.* Physical Review A, 54(3), 1838–1843.

[20] Horodecki, R., Horodecki, P., Horodecki, M., & Horodecki, K. (2009). *Quantum Entanglement* Reviews of Modern Physics, 81(2), 865–942.

[21] Jaynes, E. T. (1957). *Information theory and statistical mechanics. II.* Physical Review, 108(2), 171–190.

[22] Kahle, T. (2010). *Neighborliness of marginal polytopes.* Contributions to Algebra and Geometry, 51(1), 45–56.

[23] Kahle, T., Olbrich, E., Jost, J., & Ay, N. (2009). *Complexity measures from interaction structures.* Physical Review E, 79(2), 026201.

[24] Kahle, T., Wenzel, W., & Ay, N. (2009). *Hierarchical models, marginal polytopes, and linear codes.* Kybernetika, 45(2), 189–207.

[25] Kay, A. (2012). *Using Separable Bell-Diagonal States to Distribute Entanglement.* Physical Review Letters, 109(8), 080503.

[26] Kempe, J., Kitaev, A., & Regev, O. (2006). *The complexity of the local Hamiltonian problem.* SIAM Journal on Computing, 35(5), 1070–1097.

[27] Lang, M. D., & Caves, C. M. (2010). *Quantum discord and the geometry of Bell-diagonal states.* Physical Review Letters, 105(15), 150501.

[28] Lauritzen, S. L. (1996). *Graphical models.* Oxford University Press.
[29] Modi, K., Paterek, T., Son, W., Vedral, V., & Williamson, M. (2010). *Unified view of quantum and classical correlations.* Physical Review Letters, 104(8), 080501.
[30] Niekamp, S., Galla, T., Kleinmann, M., & Gühne, O. (2013). *Computing complexity measures for quantum states based on exponential families.* Journal of Physics A: Mathematical and Theoretical, 46(12), 125301.
[31] Nielsen, M. A., & Chuang, I. L. (2010). *Quantum computation and quantum information.* Cambridge University Press.
[32] Nielsen, M. A., & Kempe, J. (2001). *Separable states are more disordered globally than locally.* Physical Review Letters, 86(22), 5184–5187.
[33] Petz, D. (1994). *Geometry of canonical correlation on the state space of a quantum system.* Journal of Mathematical Physics, 35(2), 780–795.
[34] Rauh, J. (2011). *Finding the maximizers of the information divergence from an exponential family.* IEEE Trans. Inf. Theory, 57(6), 3236–3247.
[35] Rudolph, O. (2004). *On extremal quantum states of composite systems with fixed marginals.* Journal of Mathematical Physics, 45(11), 4035–4041.
[36] Ruelle, D. (1999). *Statistical Mechanics: Rigorous Results.* World Scientific.
[37] Shirokov, M. (2006). *Entropy characteristics of subsets of states. I.* Izvestiya: Mathematics, 70(6), 1265–1292.
[38] Vedral, V., Plenio, M. B., Rippin, M. A., & Knight, P. L. (1997). *Quantifying Entanglement.* Physical Review Letters, 78(12), 2275–2279.
[39] Von Neumann, J. (1927). *Thermodynamik quantenmechanischer Gesamtheiten.* Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse, 1927, 273–291.
[40] Weis, S. (2010). *Exponential families with incompatible statistics and their entropy distance.* Dissertation, Friedrich-Alexander-Universität Erlangen-Nürnberg.
[41] Weis, S. (2011). *Quantum convex support.* Linear Algebra and its Applications, 435(12), 3168–3188. (2012) *Correction.* ibid., 436(1), xvi.
[42] Weis, S. (2014). *Information Topologies on Non-Commutative State Spaces.* Journal of Convex Analysis, 21(2), 339–399.
[43] Weis, S., & Knauf, A. (2012). *Entropy distance: New quantum phenomena.* Journal of Mathematical Physics, 53(10), 102206.
[44] Yapage, N. and Nagaoka, H. (2008). *An information geometrical approach to the mean-field approximation for quantum Ising spin models*, J Phys A-Math Theor, 41(6), 065005.

E-mail address: maths@weis-stephan.de, knauf@math.fau.de,
E-mail address: nay@mis.mpg.de, zhaomingjingde@126.com

[1]Max-Planck-Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany

[2]Department of Mathematics, Friedrich-Alexander-University Erlangen-Nuremberg, Cauerstr. 11, D-91058 Erlangen, Germany

[3]Department of Mathematics and Computer Science, Leipzig University, PF 10 09 20, D-04009 Leipzig, Germany

[4]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA

[5]Department of Mathematics, School of Science, Beijing Information Science and Technology University, 100192, Beijing, China