The Annals of Probability
2002, Vol. 30, No. 1, 416–436

AN INFORMATION-GEOMETRIC APPROACH TO A THEORY OF PRAGMATIC STRUCTURING

BY NIHAT AY

Max-Planck-Institute for Mathematics in the Sciences

Within the framework of information geometry, the interaction among units of a stochastic system is quantified in terms of the Kullback–Leibler divergence of the underlying joint probability distribution from an appropriate exponential family. In the present paper, the main example for such a family is given by the set of all factorizable random fields. Motivated by this example, the locally farthest points from an arbitrary exponential family $\mathcal{E}$ are studied. In the corresponding dynamical setting, such points can be generated by the *structuring process* with respect to $\mathcal{E}$ as a repelling set. The main results concern the low complexity of such distributions which can be controlled by the dimension of $\mathcal{E}$.

1. Introduction.

1.1. *The motivation of the approach.* In the field of *neural networks*, so-called *infomax principles* like the principle of "maximum information preservation" by Linsker [20] are formulated to derive learning rules that improve the information processing properties of neural systems (see [12]). These principles, which are based on information-theoretic measures, are intended to describe the mechanism of learning in the brain. There, the starting point is a low-dimensional and biophysiologically motivated parametrization of the neural system, which need not necessarily be compatible with the given optimization principle. In contrast to this, we establish theoretical results about the low complexity of optimal solutions for the optimization problem of frequently used measures like the *mutual information* in an unconstrained and more theoretical setting. In the present paper, we do not comment on applications to modeling neural networks. This is intended to be done in a further step, where the results can be used for the characterization of "good" parameter sets that, on the one hand, are compatible with the underlying optimization and, on the other hand, are biologically motivated.

1.2. *An illustration of the main example.* Consider the example of two binary units with the state sets $\Omega_1 = \Omega_2 = \{0, 1\}$. The configuration set of the system is given by the product $\{0, 1\}^2$. The set $\bar{\mathcal{P}}(\{0, 1\}^2)$ of all probability distributions on that product is a three-dimensional simplex with the four extreme points $\delta_{(\omega_1,\omega_2)}$,

Received September 2000; revised March 2001.
AMS 2000 subject classifications. 62H20, 92B20, 62B05, 53B05.
Key words and phrases. Information geometry, Kullback–Leibler divergence, mutual information, infomax principle, stochastic interaction, exponential family.

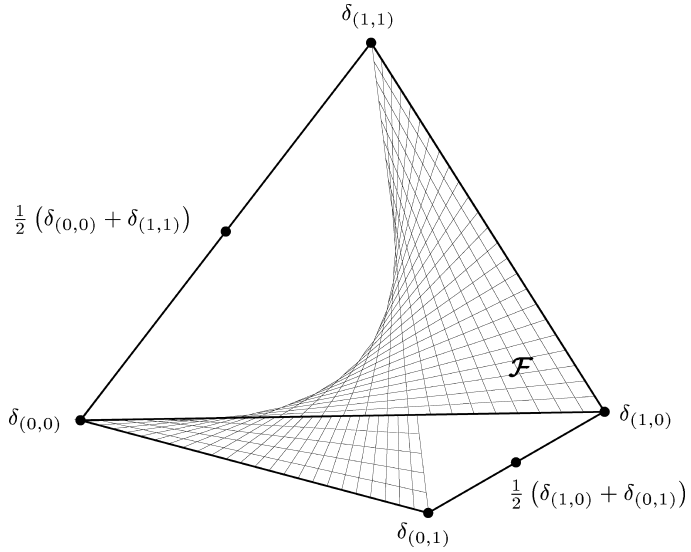

FIG. 1. *The set $\mathcal{F}$ of factorizable distributions on* $\{0,1\}^2$.

$\omega_1, \omega_2 \in \{0,1\}$ (Dirac distributions). The two units are independent with respect to $p \in \bar{\mathcal{P}}(\{0,1\}^2)$ if p is equal to the tensor product of the marginal distributions p_1 and p_2: $p = p_1 \otimes p_2$. The set of all strictly positive and in such a way factorizable distributions is a two-dimensional manifold (exponential family) $\mathcal{F}$ embedded in the simplex $\bar{\mathcal{P}}(\{0,1\}^2)$; see Figure 1.

With the *Kullback–Leibler divergence* D on $\bar{\mathcal{P}}(\{0,1\}^2)$, the dependence of the two units can be quantified by

$$\begin{aligned}\text{degree of } p\text{-dependence} &:= \text{“distance” of } p \text{ from } \mathcal{F}\\ &= \inf_{q\in\mathcal{F}} D(p\|q).\end{aligned}$$

This quantity is nothing but the mutual information of the two units with respect to p. In the present paper, we will focus on stochastic systems with the highest degree of dependence. In the example of two binary units, these are given by the distributions

$$\tfrac{1}{2}(\delta_{(0,0)}+\delta_{(1,1)}) \quad \text{and} \quad \tfrac{1}{2}(\delta_{(1,0)}+\delta_{(0,1)}) \quad \text{(see Figure 1).}$$

1.3. *The results.* Motivated by the example in Section 1.2, the farthest points from an arbitrary exponential family $\mathcal{E}$ in the set of all probability distributions are studied in a general setting by using the framework of *information geometry* for discrete probability spaces ([1], [5] and [11]). In particular, generalizations of the example of two binary units are discussed.

The results concern the low complexity of optimal distributions, which can be controlled by the dimension d of the underlying exponential family $\mathcal{E}$ (Corollary 3.4). As an important consequence of this, the existence of an exponential family $\mathcal{E}^*$ with dimension less than or equal to $\frac{1}{2}(d^2 + 7d + 4)$ that captures all points with locally maximal distance from $\mathcal{E}$ is established (Theorem 3.5). In particular, for the example of N binary units, $N \geq 8$, there is an exponential family with dimension less than or equal to N^2 that captures all distributions with optimal dependence of the units.

A translation of the setting into a dynamical version is given by the definition of *structuring processes* with respect to exponential families as repelling sets (Theorem 3.12). The stable limit points of such a process play the role of distributions with largest distance from the underlying exponential family. In the context of neural networks, structuring is related to learning that is induced by the infomax principles mentioned in Section 1.1.

2. Notation and preliminaries. In the following, Ω denotes a nonempty and finite set. With the usual addition and scalar multiplication, the set $\mathbb{R}^\Omega$ of all functions $\Omega \to \mathbb{R}$ becomes a real vector space. In $\mathbb{R}^\Omega$ we have the canonical basis

$$e_\omega\colon\ \omega' \mapsto e_\omega(\omega') := \begin{cases} 1, & \text{if } \omega' = \omega, \\ 0, & \text{otherwise,} \end{cases} \qquad \omega \in \Omega,$$

which induces the norm $\|x\| = (\sum_{\omega\in\Omega} x(\omega)^2)^{1/2}$.

The (closed) simplex

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

is a convex and compact subset of $\mathbb{R}^\Omega$ with the extreme points e_ω, $\omega \in \Omega$. Its elements are the probability measures on Ω. The extreme points correspond to the Dirac measures, and the centroid $c \in \bar{\mathcal{P}}(\Omega)$ with $c(\omega) := 1/|\Omega|$ for all $\omega \in \Omega$ is the equally distributed normed measure. For all $x \in \mathbb{R}^\Omega$, $\operatorname{supp} x := \{\omega \in \Omega : x(\omega) \neq 0\}$ denotes the support set of x. Every nonempty subset Σ of Ω induces the corresponding (open) face

$$\mathcal{P}(\Sigma) := \big\{ p \in \bar{\mathcal{P}}(\Omega) : \operatorname{supp} p = \Sigma \big\}$$

of $\bar{\mathcal{P}}(\Omega)$. Obviously, the closed simplex is the disjoint union

$$\bar{\mathcal{P}}(\Omega) = \biguplus_{\varnothing\neq\Sigma\subset\Omega} \mathcal{P}(\Sigma)$$

of the faces, and every element $p \in \bar{\mathcal{P}}(\Omega)$ is contained in $\mathcal{P}(\operatorname{supp} p)$.

Following the information-geometric description of finite probability spaces, each open face $\mathcal{P}(\Sigma)$ can be considered as a differentiable submanifold of $\mathbb{R}^\Omega$

with dimension $d := |\Sigma| - 1$ and the basis-point independent tangent space

$$\mathrm{T}(\Sigma) := \left\{ x \in \mathbb{R}^{\Omega} : \operatorname{supp} x = \Sigma,\ \sum_{\omega\in\Sigma} x(\omega) = 0 \right\}.$$

With the *Fisher metric* $\langle\cdot,\cdot\rangle_p$: $\mathrm{T}(\Sigma) \times \mathrm{T}(\Sigma) \to \mathbb{R}$ in $p \in \mathcal{P}(\Sigma)$ defined by

$$(x, y) \mapsto \langle x, y\rangle_p := \sum_{\omega\in\Sigma} \frac{1}{p(\omega)} x(\omega) y(\omega),$$

$\mathcal{P}(\Sigma)$ becomes a Riemannian manifold. The most important additional structure studied in information geometry is given by a pair of dual affine connections on the manifold. Application of such a dual structure to the present situation leads to the notion of (-1)- and $(+1)$-geodesics: Each two points $p, q \in \mathcal{P}(\Sigma)$ can be connected by the geodesics $\gamma^{(\alpha)} = (\gamma_\omega^{(\alpha)})_{\omega\in\Sigma}$: $[0, 1] \to \mathcal{P}(\Sigma)$, $\alpha \in \{-1, +1\}$, with

$$\gamma_\omega^{(-1)}(t) := (1 - t)\, p(\omega) + t q(\omega) \quad \text{and} \quad \gamma_\omega^{(+1)}(t) := r(t) p(\omega)^{1-t} q(\omega)^t.$$

Here, $r(t)$ denotes the normalization factor.

A submanifold $\mathcal{E}$ of $\mathcal{P}(\Omega)$ is called an *exponential family* if there exist a point $p_0 \in \mathcal{P}(\Omega)$ and vectors $v_1, \ldots, v_d \in \mathbb{R}^{\Omega}$ such that it is the image of the map

$$\mathbb{R}^d \to \mathcal{P}(\Omega), \qquad (\theta_1, \ldots, \theta_d) \mapsto \sum_{\omega\in\Omega} \frac{p_0(\omega) \exp\bigl(\sum_{i=1}^d \theta_i v_i(\omega)\bigr)}{\sum_{\omega'\in\Omega} p_0(\omega') \exp\bigl(\sum_{i=1}^d \theta_i v_i(\omega')\bigr)} e_\omega.$$

In this case, the $(+1)$-geodesically convex manifold $\mathcal{E}$ is said to be generated by p_0 and $v_1, \ldots, v_d$. One has $\dim \mathcal{E} \le d$, where the equality holds if and only if the vectors $\{v_1, \ldots, v_d, 1\}$ are linearly independent (1 denotes the "constant" vector with entries equal to 1).

A general projection theorem by Amari ([1], Theorem 3.9, page 91) implies the following:

Let $\mathcal{E}$ be an exponential family and $p \in \mathcal{P}(\Omega)$. Then there exists at most one point $p' \in \mathcal{E}$ such that the (-1)-geodesic connecting p and p' intersects $\mathcal{E}$ orthogonally with respect to the Fisher metric.

Such a point p' is called a (-1)-*projection* of p onto $\mathcal{E}$ and can be characterized by the *Kullback–Leibler divergence* D: $\bar{\mathcal{P}}(\Omega) \times \bar{\mathcal{P}}(\Omega) \to \overline{\mathbb{R}}$,

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

We define the "distance" $D_{\mathcal{E}}$: $\bar{\mathcal{P}}(\Omega) \to \mathbb{R}_+$ from $\mathcal{E}$ by

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

This is a continuous function (Lemma 4.2). It was proven by Amari that a point $p' \in \mathcal{E}$ is the (-1)-projection of p onto $\mathcal{E}$ if and only if it satisfies the minimizing property $D_{\mathcal{E}}(p) = D(p\|p')$ ([1], Theorem 3.8, page 90). With $\operatorname{dom}\mathcal{E}$ we denote the set of all points in $\bar{\mathcal{P}}(\Omega)$ for which there exist such distance minimizing points in $\mathcal{E}$ and define the corresponding projection $\pi_{\mathcal{E}}\colon \operatorname{dom}\mathcal{E} \to \mathcal{E}$.

MAIN EXAMPLE, PART 1. Let $\Omega_1, \ldots, \Omega_N$ be nonempty and finite sets. Consider the *tensorial map* $\mathcal{P}(\Omega_1) \times \cdots \times \mathcal{P}(\Omega_N) \to \mathcal{P}(\Omega_1 \times \cdots \times \Omega_N)$,

$$(p_1, \ldots, p_N) \mapsto p_1 \otimes \cdots \otimes p_N := \sum_{\substack{(\omega_1,\ldots,\omega_N)\\ \in \Omega_1 \times \cdots \times \Omega_N}} p_1(\omega_1) \cdots p_N(\omega_N) e_{(\omega_1,\ldots,\omega_N)}.$$

The image $\mathcal{F} := \{p_1 \otimes \cdots \otimes p_N : p_i \in \mathcal{P}(\Omega_i),\ 1 \le i \le N\}$ of this map, which consists of all factorizable and strictly positive probability distributions, is an exponential family in $\mathcal{P}(\Omega_1 \times \cdots \times \Omega_N)$ with $\dim \mathcal{F} = (|\Omega_1| - 1) + \cdots + (|\Omega_N| - 1)$. For the particular case of N binary units, that is, $|\Omega_i| = 2$ for all i, the dimension of $\mathcal{F}$ is equal to N. The following statements are well known (see [3]):

$$\pi_{\mathcal{F}}(p) = p_1 \otimes \cdots \otimes p_N, \qquad D_{\mathcal{F}}(p) = \sum_{i=1}^{N} H(p_i) - H(p). \tag{2.1}$$

Here, p_i denotes the ith marginal distribution of p and H the Shannon entropy [28].

3. Results and applications.

NOTE. All proofs are given in Section 4.

3.1. *The main results in the nondynamical setting.*

PROPOSITION 3.1. *For an exponential family $\mathcal{E}$ in $\mathcal{P}(\Omega)$ and a point $p \in \operatorname{dom}\mathcal{E}$, the gradient of $D_{\mathcal{E}}$ in p with respect to the Fisher metric is given by*

$$\operatorname{grad} D_{\mathcal{E}}(p) = \sum_{\omega \in \operatorname{supp} p} p(\omega)\left(\ln \frac{p(\omega)}{\pi_{\mathcal{E}}(p)(\omega)} - D_{\mathcal{E}}(p)\right) e_\omega. \tag{3.2}$$

In particular, if the gradient vanishes in p, then we have

$$p(\omega) = e^{D_{\mathcal{E}}(p)} \pi_{\mathcal{E}}(p)(\omega),\ \omega \in \operatorname{supp} p \quad \text{and} \quad D_{\mathcal{E}}(p) = -\ln \pi_{\mathcal{E}}(p)(\operatorname{supp} p). \tag{3.3}$$

PROPOSITION 3.2. *Let $\mathcal{E}$ be an exponential family in $\mathcal{P}(\Omega)$ and $p \in \operatorname{dom}\mathcal{E}$. If $D_{\mathcal{E}}$ attains a locally maximal value in p, then the cardinality of the support set of p can be estimated by $|\operatorname{supp} p| \le \dim \mathcal{E} + 1$.*

REMARKS 3.3. (i) For the case $\mathcal{E} := \mathcal{P}(\Omega)$, $D_{\mathcal{E}}$ vanishes identically and the statements in Propositions 3.1 and 3.2 are trivially satisfied for all points in $\mathcal{P}(\Omega)$.

(ii) If the exponential family consists of a single point, Proposition 3.2 implies that we have only Dirac measures as locally maximal points.

An immediate consequence of Propositions 3.1 and 3.2 is the following statement about the structure of distributions with locally largest distance from the underlying exponential family.

COROLLARY 3.4. *Let $\mathcal{E}$ be a d-dimensional exponential family generated by p_0 and $v_1, \dots, v_d$. If $D_{\mathcal{E}}$ attains a locally maximal value in $p \in \operatorname{dom}\mathcal{E}$, then there exist real numbers $r_1, \dots, r_d$ and a set $\Sigma \subset \Omega$ with $|\Sigma| \leq d+1$ such that the following representation of p holds*:

$$p(\omega) = \begin{cases} \dfrac{p_0(\omega)\exp\left(\sum_{i=1}^{d} r_i v_i(\omega)\right)}{\sum_{\omega' \in \Sigma} p_0(\omega')\exp\left(\sum_{i=1}^{d} r_i v_i(\omega')\right)}, & \omega \in \Sigma, \\ 0, & \omega \notin \Sigma. \end{cases}$$

Here, the numbers $r_1, \dots, r_d$ and the set Σ are unique.

The "exponential" structure of this representation indicates the possibility of capturing all optimal distributions with respect to $D_{\mathcal{E}}$ by an exponential family $\mathcal{E}^*$ with low dimension. This is guaranteed by the following.

THEOREM 3.5. *Let $\mathcal{E}$ be a d-dimensional exponential family. Then there exists an exponential family $\mathcal{E}^* \supset \mathcal{E}$ with dimension less than or equal to $\frac{1}{2}(d^2+7d+4)$ such that the topological closure of $\mathcal{E}^*$ contains all locally maximal points in* $\operatorname{dom}\mathcal{E}$ *with respect to $D_{\mathcal{E}}$.*

REMARK 3.6. By choosing $\mathcal{E}$ to be the exponential family that consists of the centroid of $\mathcal{P}(\Omega)$, $|\Omega| \geq 3$, Theorem 3.5 implies that there exists a two-dimensional exponential family $\mathcal{E}^*$ in $\mathcal{P}(\Omega)$ such that all extreme points of the simplex $\bar{\mathcal{P}}(\Omega)$ can be approximated by $\mathcal{E}^*$. To construct such a family, choose an arbitrary numbering $\varphi\colon \Omega \to \{1, \dots, |\Omega|\} \subset \mathbb{R}$ of the set Ω and define $\mathcal{E}^*$ to be the two-dimensional exponential family generated by the centroid and the vectors φ and φ^2. For $1 \leq k \leq |\Omega|$ and $\beta_n \uparrow \infty$, we have

$$\lim_{n\to\infty} \frac{\exp(-\beta_n(\varphi(\omega) - \phi(\sigma))^2)}{\sum_{\omega' \in \Omega}\exp(-\beta_n(\varphi(\omega') - \phi(\sigma))^2)} = \delta_\sigma(\omega) \qquad \text{for all } \sigma, \omega \in \Omega.$$

Thus, $\mathcal{E}^*$ approximates all Dirac measures.

This is the finite-dimensional counterpart of the fact that all Dirac measures on $(\mathbb{R}, \mathscr{B}^1)$ can be approximated by normal distributions (two-dimensional exponential family) in the sense of weak convergence.

3.2. *Some applications.* The examples considered in the present paper are induced by a special kind of exponential family. Let $\mathscr{A}$ be a subset of the power set $\mathscr{P}(\Omega)$ of Ω. With the characteristic functions $\mathbb{I}_A\colon \Omega \to \mathbb{R}$, $\mathbb{I}_A(\omega) = 1$ if $\omega \in A$ and $\mathbb{I}_A(\omega) = 0$ if $\omega \notin A$, we define $\mathcal{E}_{\mathscr{A}}$ to be the exponential family generated by the centroid of $\mathcal{P}(\Omega)$ and the functions $\mathbb{I}_A$, $A \in \mathscr{A}$. The following statement gives a description of the support set of an element that is projectable on $\mathcal{E}_{\mathscr{A}}$:

LEMMA 3.7. *Let $\mathscr{A}$ be a subset of $\mathscr{P}(\Omega)$ and $p \in \operatorname{dom}\mathcal{E}_{\mathscr{A}}$. Then for every nonempty set $A \subset \Omega$ with $A \in \mathscr{A}$ or $\Omega \setminus A \in \mathscr{A}$ the intersection* $\operatorname{supp} p \cap A$ *is also nonempty.*

MAIN EXAMPLE, PART 2. For each $i \in \{1, \dots, N\}$, consider the partition

$$\mathscr{A}_i := \{\Omega_1 \times \cdots \times \{\omega_i\} \times \cdots \times \Omega_N : \omega_i \in \Omega_i\}$$

of $\Omega_1 \times \cdots \times \Omega_N$. The exponential family $\mathcal{F}$ of the factorizable distributions in $\mathcal{P}(\Omega_1 \times \cdots \times \Omega_N)$ is induced by

$$\mathscr{A} := \bigcup_{i=1}^{N} \mathscr{A}_i.$$

Thus, we have $\mathcal{E}_{\mathscr{A}} = \mathcal{F}$. Let p be a point in $\operatorname{dom}\mathcal{F} \subset \bar{\mathcal{P}}(\Omega_1 \times \cdots \times \Omega_N)$. Lemma 3.7 implies $|\operatorname{supp} p| \geq |\mathscr{A}_i| = |\Omega_i|$ for all i. Therefore, we obtain

$$|\operatorname{supp} p| \geq \max_{1 \leq i \leq N} |\Omega_i|.$$

If $D_{\mathcal{F}}$ attains a locally maximal value in p, then with (2.1) and (3.3) we get

$$p(\omega_1, \dots, \omega_N) = \exp\left(\sum_{i=1}^{N} H(p_i) - H(p)\right) p_1(\omega_1) \cdots p_N(\omega_N)$$

for all $(\omega_1, \dots, \omega_N) \in \operatorname{supp} p$. According to Proposition 3.2, the cardinality of the support set can be estimated by

$$\max_{1 \leq i \leq N} |\Omega_i| \leq |\operatorname{supp} p| \leq 1 - N + \sum_{i=1}^{N} |\Omega_i|.$$

With $\eta := \max_{1 \leq i \leq N}(|\Omega_i| - 1)$, we obtain

$$\eta \leq |\operatorname{supp} p| - 1 \leq N\eta.$$

Finally, for $N \geq 8$, Theorem 3.5 guarantees the existence of an exponential family $\mathcal{F}^*$ with dimension $\leq (N\eta)^2$ that approximates all locally maximal points with respect to $D_{\mathcal{F}}$. These are the points with optimal dependence of the N units.

GENERALIZATION OF THE MAIN EXAMPLE. For every subset $J \subset I := \{1, \ldots, N\}$, $J \neq \varnothing$, consider the restriction

$$\mathrm{rest}_J \colon \prod_{i \in I} \Omega_i \to \prod_{i \in J} \Omega_i, \qquad (\omega_i)_{i \in I} \mapsto (\omega_i)_{i \in J}$$

and

$$\mathscr{A}_J := \left\{ \mathrm{rest}_J^{-1}(\{\omega_J\}) : \omega_J \in \prod_{i \in J} \Omega_i \right\}.$$

For a fixed n, $1 \leq n \leq N$, we define

$$\mathscr{A}^{(n)} := \bigcup_{\substack{J \subset I \\ 1 \leq |J| \leq n}} \mathscr{A}_J \quad \text{and} \quad \mathcal{F}^{(n)} := \mathcal{E}_{\mathscr{A}^{(n)}}.$$

Each $\mathcal{F}^{(n)}$ represents only intrinsic dependencies up to order n. With $\mathcal{F}^{(0)} := \{c\}$, we have the hierarchy

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

and the equations $\mathcal{F}^{(1)} = \mathcal{F}$, $\mathcal{F}^{(N)} = \mathcal{P}(\Omega)$ hold.

For simplicity, in the following we consider only the binary case $\Omega_i = \{0, 1\}$ for all i. In that case, the exponential family $\mathcal{F}^{(n)}$ consists of all strictly positive probability distributions in $\mathcal{P}(\{0, 1\}^I)$ for which there exist real numbers θ_J, $J \subset I$, $|J| \leq n$, such that for all $\sigma = (\sigma_i)_{i \in I} \in \{0, 1\}^I$ the equality

$$\ln p(\sigma) = \sum_{\substack{J \subset I \\ |J| \leq n}} \theta_J \cdot \prod_{i \in J} \sigma_i$$

holds ($\prod_{i \in \varnothing} \sigma_i := 1$). Furthermore, one has

$$\dim \mathcal{F}^{(n)} = \sum_{i=1}^{n} \binom{N}{i}$$

(for these statements see [3] and [21]). Now we apply Corollary 3.4 and Lemma 3.7 to a point $p \in \mathrm{dom}\, \mathcal{F}^{(n)}$ in which $D_{\mathcal{F}^{(n)}}$ attains a locally maximal value: There exist real numbers θ_J, $J \subset I$, $1 \leq |J| \leq n$, and a set $\Sigma \subset \Omega$ with

$$2^n = \sum_{i=0}^{n} \binom{n}{i} \leq |\Sigma| \leq \sum_{i=0}^{n} \binom{N}{i} \leq \min\{(N+1)^n, 2^N\}$$

and

$$p(\sigma) = \frac{\exp\left(\sum_{J \subset I, 1 \leq |J| \leq n} \theta_J \cdot \prod_{i \in J} \sigma_i\right)}{\sum_{\sigma' \in \Sigma} \exp\left(\sum_{J \subset I, 1 \leq |J| \leq n} \theta_J \cdot \prod_{i \in J} \sigma'_i\right)} \mathbb{I}_\Sigma(\sigma).$$

3.3. *Structuring fields and the dynamical setting.*

DEFINITION 3.8. We call the vector field $\Psi_{\mathcal{E}}\colon \mathcal{P}(\Omega) \to \mathrm{T}(\Omega)$ defined by

$$p \mapsto \Psi_{\mathcal{E}}(p) := \operatorname{grad} D_{\mathcal{E}}(p)$$

the *structuring field* with respect to $\mathcal{E}$.

The most trivial example is given by $\mathcal{E} := \mathcal{P}(\Omega)$. In that case the structuring field vanishes identically and there is no "motion." The complementary situation where $\mathcal{E}$ consists of only one point is discussed in Example 3.13.

PROPOSITION 3.9. *Let $\mathcal{E}$ be an exponential family in $\mathcal{P}(\Omega)$. Then for every initial point $p_0 \in \mathcal{P}(\Omega)$ there exists a unique maximal solution $\gamma\colon I \to \mathcal{P}(\Omega)$ for the problem*

$$\frac{d\gamma}{dt} = \Psi_{\mathcal{E}}(\gamma), \qquad \gamma(0) = p_0. \tag{3.4}$$

If $\lim_{t\to\sup I} \gamma(t)$ exists and is projectable, then $\sup I = \infty$.

To translate the results stated in Section 3.1 into the dynamical setting, we define the following:

DEFINITION 3.10. A point $p \in \bar{\mathcal{P}}(\Omega)$ is a (*positive*) *limit point* with respect to $\Psi_{\mathcal{E}}$ iff there exists a solution $\gamma\colon I \to \mathcal{P}(\Omega)$ for $\Psi_{\mathcal{E}}$ that converges to p: $\lim_{t\to\sup I} \gamma(t) = p$. The limit point p is *stable* iff there exists an open neighborhood U of p in $\bar{\mathcal{P}}(\Omega)$ such that for every point $p_0 \in U \cap \mathcal{P}(\Omega)$ there exists a solution with initial point p_0 that converges to p.

REMARK 3.11. The correspondence between stable limit points and locally maximal points with respect to $D_{\mathcal{E}}$ is not one to one. The property to be a locally maximal point does not imply the stability in the sense of Definition 3.10.

THEOREM 3.12. *Let $\mathcal{E}$ be an exponential family in $\mathcal{P}(\Omega)$ and $p \in \operatorname{dom}\mathcal{E}$ a limit point with respect to the structuring field $\Psi_{\mathcal{E}}$. Then the statements* (3.3) *in Proposition* 3.1 *are valid. If the limit point p is stable, the cardinality of its support set can be estimated as in Proposition* 3.2.

EXAMPLE 3.13. If the exponential family $\mathcal{E}$ consists of only one point $q \in \mathcal{P}(\Omega)$, then the projection is given by $\pi_{\mathcal{E}}(p) = q$ for all $p \in \mathcal{P}(\Omega)$. We have the structuring field

$$p \mapsto \operatorname{grad} D_{\mathcal{E}}(p) = \sum_{\omega\in\Omega} p(\omega)\left(\ln\frac{p(\omega)}{q(\omega)} - D(p\|q)\right)e_\omega.$$

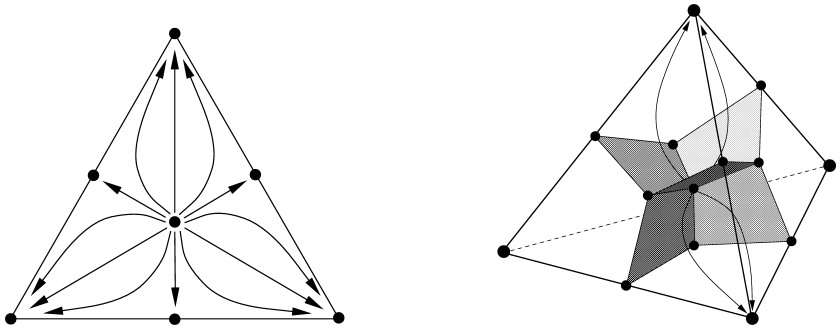

FIG. 2. *Structuring field with respect to the centroid in dimensions* 2 *and* 3.

With an initial point $p_0 \in \mathcal{P}(\Omega)$, we obtain the following solution $\gamma = (\gamma_\omega)_{\omega\in\Omega}$: $\mathbb{R} \to \mathcal{P}(\Omega)$ for the problem (3.4):

$$\gamma_\omega(t) = \frac{q(\omega)^{1-e^t} p_0(\omega)^{e^t}}{\sum_{\omega'\in\Omega} q(\omega')^{1-e^t} p_0(\omega')^{e^t}}. \tag{3.5}$$

The trajectory of γ is a $(+1)$-geodesic going through p_0. Furthermore, we have, for all ω,

$$\lim_{t\to-\infty} \gamma_\omega(t) = q(\omega)$$

and

$$\lim_{t\to+\infty} \gamma_\omega(t) = \frac{q(\omega)}{q(M)}\mathbb{I}_M(\omega) \qquad \text{with } M := \arg\max_{\omega'} \frac{p_0(\omega')}{q(\omega')}.$$

For a generic p_0, M has only one element ω and the orbit converges to the Dirac measure that is concentrated in ω. Figure 2 illustrates the flow in $\mathcal{P}(\Omega)$ for $|\Omega| = 3$ and $|\Omega| = 4$ with respect to the centroid as repelling set.

The trajectories are related to those appearing in statistical physics by variation of the *inverse temperature* β. There, one considers distributions of the form

$$p_\beta(\omega) = \frac{e^{-\beta E(\omega)}}{\sum_{\omega'\in\Omega} e^{-\beta E(\omega')}},$$

where E denotes the *energy function*. Setting $\beta := e^t$ and $E := -\ln p_0$, one obtains a special case of (3.5).

MAIN EXAMPLE, PART 3 (Some computer simulations). With (2.1) we get the structuring field

$$p \mapsto \Psi_{\mathcal{F}}(p) = \sum_{\substack{\omega=(\omega_1,\dots,\omega_N)\\ \in\Omega_1\times\cdots\times\Omega_N}} p(\omega)\left(\ln p(\omega) + H(p) - \sum_{i=1}^{N}\big(\ln p_i(\omega_i) + H(p_i)\big)\right)e_\omega,$$

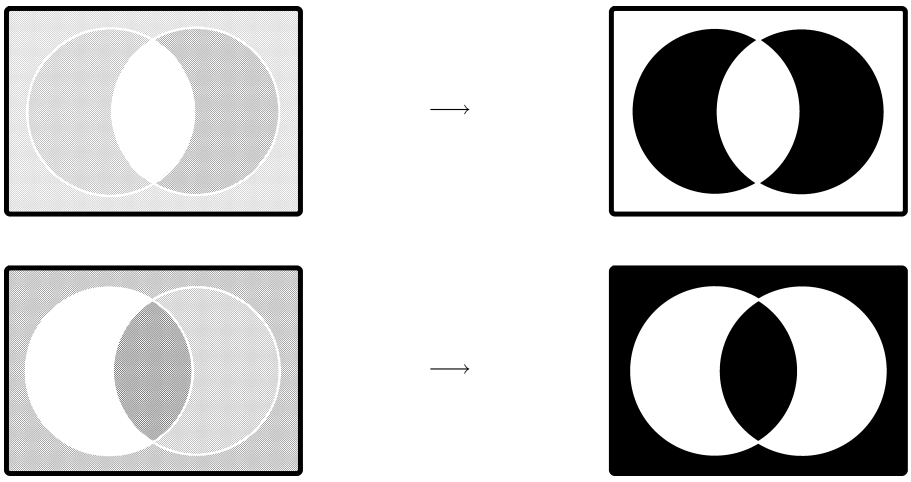

FIG. 3. *Simulations of $\Psi_{\mathcal{F}}$ for two binary units.*

with respect to the exponential family $\mathcal{F}$ of the factorizable distributions on the product set $\Omega_1 \times \cdots \times \Omega_N$. For the simulation of the corresponding process, with an initial point p_0 and a sequence $\varepsilon_n \downarrow 0$, we define the following iteration rule:

$$p^{(0)} := p_0 \quad \text{and} \quad p^{(n+1)} := r_n p^{(n)} \left(\frac{p^{(n)}}{p_1^{(n)} \otimes \cdots \otimes p_N^{(n)}} \right)^{\varepsilon_n}, \qquad n = 0, 1, 2, \ldots .$$

Here, r_n is the normalization factor at time n. This iteration, which follows the *gradient method* with respect to the $(+1)$-geodesics, has not been analyzed analytically.

In the following, the structuring process is illustrated by some computer simulations, starting with the case of binary units:

(i) In the Venn diagrams shown in Figures 3 and 4, each circle represents one unit. The interior and the exterior of such a circle are the disjoint events in the configuration set of the system corresponding to the two states of the unit. Each diagram illustrates a probability distribution on the set of all atoms, which are generated by the events of all units in the system. The gray value of an atom is proportional to the probability of the atom; "white" is the maximal probability in a given Venn diagram. The diagrams on the left-hand side are the initial distributions and the ones on the right-hand side are the limit (structured) distributions.

We start with two units (see Figure 3); this is the situation that has been discussed in Section 1.2. The stable limit distributions are $\frac{1}{2}(\delta_{(0,0)} + \delta_{(1,1)})$ and $\frac{1}{2}(\delta_{(1,0)} + \delta_{(0,1)})$ (see Figure 1). Figure 4 gives some examples for three binary units.

(ii) Now consider two units with the state sets $\Omega_1 = \{1, 2, \ldots, m\}$ and $\Omega_2 = \{1, 2, \ldots, n\}$. Every $(m \times n)$-field in the simulations shown in Figure 5 represents a

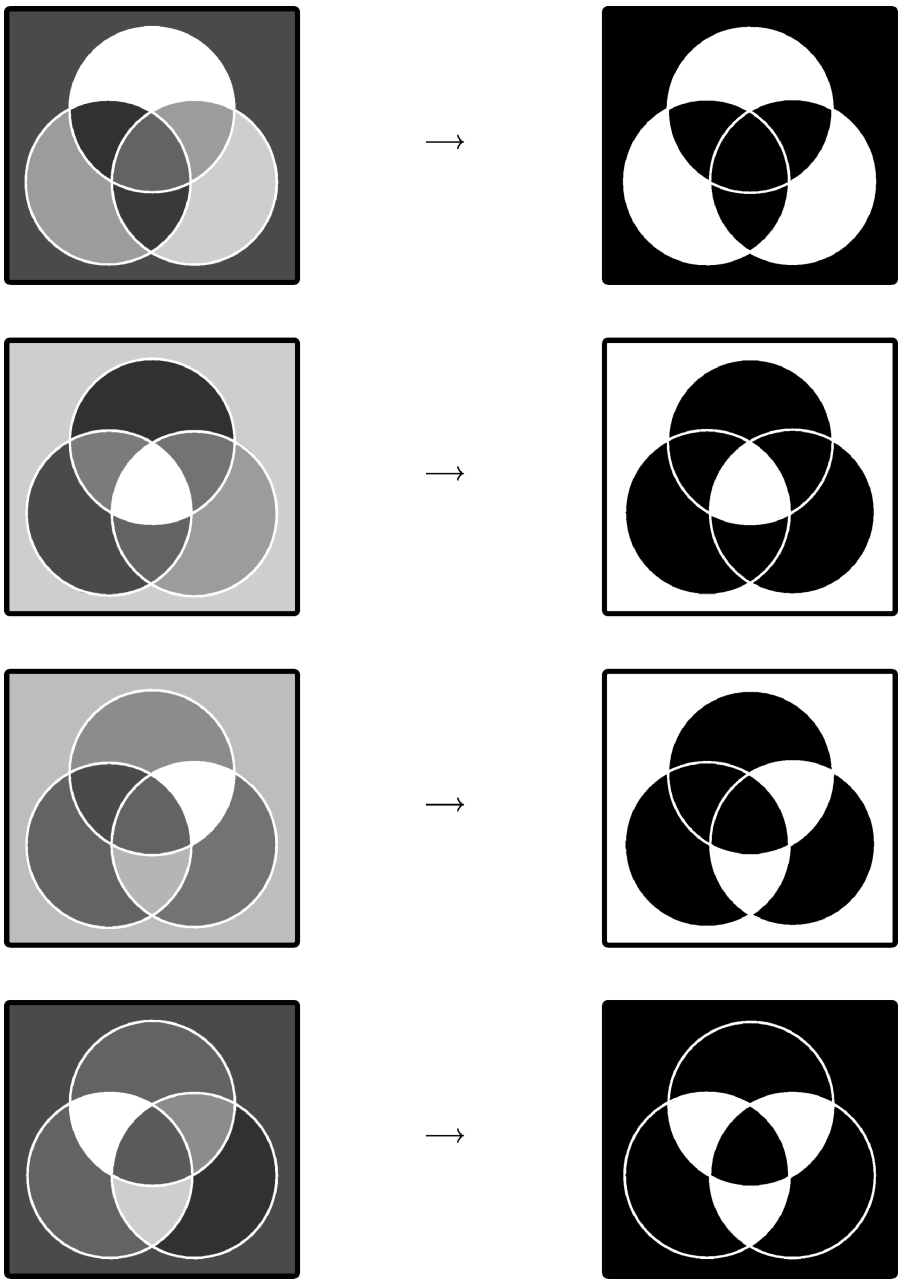

FIG. 4. *Simulations of $\Psi_{\mathcal{F}}$ for three binary units.*

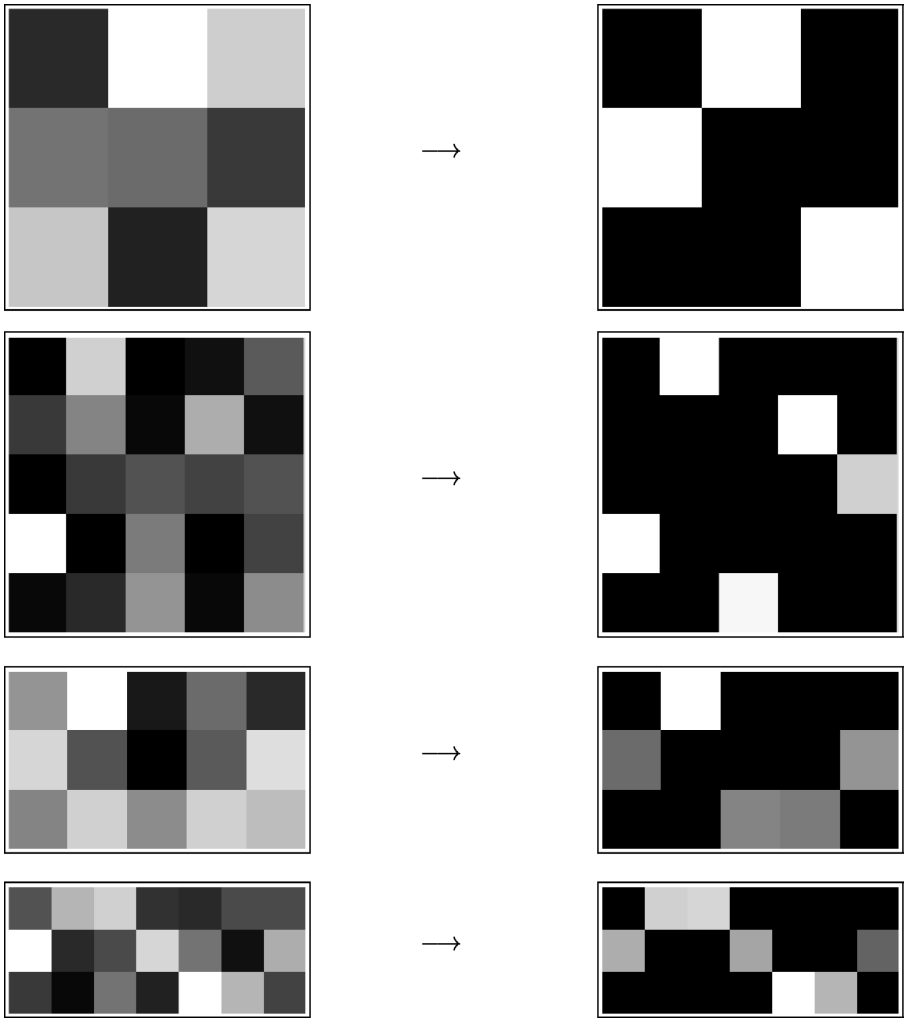

FIG. 5. *Simulations of $\Psi_{\mathcal{F}}$ for some cardinalities $m, n \geq 3$.*

probability distribution on the configuration set $\Omega_1 \times \Omega_2$. The horizontal direction of each field corresponds to the elements of Ω_1 and the vertical to the ones of Ω_2. The gray value of an event $(i, j) \in \Omega_1 \times \Omega_2$ is proportional to its probability; "white" is the maximal probability in a given field.

With these examples, one can see that the process has the tendency to structure the initial distribution in such a way that the support set becomes a graph of a one-to-one mapping between the state sets of the two units. Of course, this is only possible for the case $m = n$. The situation $m > n$ is illustrated in the last two simulations in Figure 5, where the support set of each final distribution is the

graph of a surjective mapping. This is a consequence of the entropy maximization in each unit [see (2.1)].

REMARK 3.14. Within the framework of information geometry, gradient fields have been studied in [13] and [24]. In [13], although the underlying mathematical structure is more general than in the present paper, the flows correspond to our situation of a one-point exponential family as in Example 3.13. It has been proven ([13], Theorem 1) that the trajectories of such flows are in general of geodesic type.

3.4. *Some problems and comments.* (i) Dependency among stochastic units is frequently referred to as "stochastic interaction" [3]. Of course, the dynamical aspects of interaction are ignored in the present approach. A generalization of this approach to Markov processes is necessary for a better understanding of the dynamical properties of strongly interacting units.

(ii) Theorem 3.5 guarantees the existence of low-dimensional exponential families $\mathcal{E}^*$ that capture all optimal distributions with respect to $D_{\mathcal{E}}$. One has to construct such families more explicitly in order to define models for learning systems.

(iii) From the learning-theoretical point of view, statements like Theorem 3.5 are interesting for the reason that they provide a characterization of parameter sets for learning systems with high generalization ability. Although the setting of *statistical learning theory* by Vapnik and Chervonenkis [29] is different from the present one, a broad notion of generalization can be captured by its mathematical basis given in [30].

4. Proofs.

LEMMA 4.1. *Let $\mathcal{E}$ be a d-dimensional exponential family in $\mathcal{P}(\Omega)$ generated by p_0 and $v_1, \dots, v_d$. If a point p is projectable on $\mathcal{E}$, that is, $p \in \operatorname{dom}\mathcal{E}$, then there exist a neighborhood U of p in $\mathbb{R}^\Omega$ and continuously differentiable functions $l_i\colon U \to \mathbb{R}$, $i = 1, \dots, d$, such that for all $q \in U \cap \bar{\mathcal{P}}(\Omega)$ the following holds*:

$$q \in \operatorname{dom}\mathcal{E} \quad \textit{and} \quad \pi_{\mathcal{E}}(q) = \sum_{\omega\in\Omega} \frac{\exp\bigl(\sum_{i=1}^d l_i(q)v_i(\omega)\bigr)}{\sum_{\omega'\in\Omega}\exp\bigl(\sum_{i=1}^d l_i(q)v_i(\omega')\bigr)} e_\omega.$$

LEMMA 4.2. *Let $\mathcal{E}$ be an exponential family in $\mathcal{P}(\Omega)$. Then $D_{\mathcal{E}}$ is continuous on $\bar{\mathcal{P}}(\Omega)$.*

PROOF. Let p be a point in $\bar{\mathcal{P}}(\Omega)$ and $p_n \in \bar{\mathcal{P}}(\Omega)$, $n \in \mathbb{N}$, a sequence with $p_n \to p$. For all $q \in \mathcal{E} \subset \mathcal{P}(\Omega)$, we have

$$D_{\mathcal{E}}(p_n) \le D(p_n \,\|\, q). \tag{4.6}$$

The continuity of $D(\cdot \,\|\, q)$ implies the convergence

$$\lim_{n\to\infty} D(p_n \,\|\, q) = D(p \,\|\, q). \tag{4.7}$$

With (4.6) and (4.7), one has

$$\limsup_{n\to\infty} D_{\mathcal{E}}(p_n) \le \limsup_{n\to\infty} D(p_n \,\|\, q) = \lim_{n\to\infty} D(p_n \,\|\, q) = D(p \,\|\, q). \tag{4.8}$$

From the lower semicontinuity of the Kullback–Leibler divergence D (see, e.g., [9], [16] and [18]), we get the same continuity property for the map $D_{\mathcal{E}}$ (see [26], Theorem 1.17, page 16). With this, taking the infimum of the right-hand side of (4.8) leads to

$$\limsup_{n\to\infty} D_{\mathcal{E}}(p_n) \le \inf_{q\in\mathcal{E}} D(p \,\|\, q) = D_{\mathcal{E}}(p) \tag{4.9}$$

$$\le \liminf_{n\to\infty} D_{\mathcal{E}}(p_n). \tag{4.10}$$

So, the equality holds in (4.9) and (4.10) and we finally get

$$\lim_{n\to\infty} D_{\mathcal{E}}(p_n) = D_{\mathcal{E}}(p). \qquad \square$$

PROOF OF PROPOSITION 3.1. With an arbitrary numbering $\Sigma := \operatorname{supp} p = \{\omega_1, \ldots, \omega_{n+1}\}$, we consider the coordinate chart

$$\varphi\colon \left\{(\theta_1, \ldots, \theta_n) \in \mathbb{R}^n : \theta_i > 0 \text{ for all } i,\ \sum_{i=1}^{n} \theta_i < 1\right\} \to \mathcal{P}(\Sigma),$$

with

$$\theta = (\theta_1, \ldots, \theta_n) \mapsto \varphi(\theta_1, \ldots, \theta_n) := \sum_{i=1}^{n} \theta_i e_{\omega_i} + \left(1 - \sum_{i=1}^{n} \theta_i\right) e_{\omega_{n+1}}.$$

The tangent space $\mathrm{T}(\Sigma)$ is spanned by the vectors $\partial_i := \partial\varphi/\partial\theta_i = e_{\omega_i} - e_{\omega_{n+1}}$, $i = 1, \ldots, n$, and the Fisher metric in p is given by the matrix

$$g_{ij}(p) := \langle \partial_i, \partial_j \rangle_p = \frac{1}{p(\omega_{n+1})} \begin{pmatrix} \frac{p(\omega_{n+1})}{p(\omega_1)} + 1 & 1 & \cdots & 1 \\ 1 & \frac{p(\omega_{n+1})}{p(\omega_2)} + 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & \frac{p(\omega_{n+1})}{p(\omega_n)} + 1 \end{pmatrix}$$

(see [1], Example 2.4, page 31). We have the corresponding inverse matrix

$$g^{ij}(p) = \begin{pmatrix} p(\omega_1)\big(1 - p(\omega_1)\big) & -p(\omega_1)p(\omega_2) & \cdots & -p(\omega_1)p(\omega_n) \\ -p(\omega_2)p(\omega_1) & p(\omega_2)\big(1 - p(\omega_2)\big) & \cdots & -p(\omega_1)p(\omega_n) \\ \vdots & \vdots & \ddots & \vdots \\ -p(\omega_n)p(\omega_1) & -p(\omega_n)p(\omega_2) & \cdots & p(\omega_n)\big(1 - p(\omega_n)\big) \end{pmatrix}.$$

Let $\mathcal{E}$ be a d-dimensional exponential family generated by p_0 and $v_1, \dots, v_d$. According to Lemma 4.1, in a neighborhood of $\varphi^{-1}(p)$ we have the representation

$$\pi_{\mathcal{E}}\big(\varphi(\theta)\big)(\omega) = \frac{\exp\big(\sum_{i=1}^{d} l_i(\varphi(\theta))v_i(\omega)\big)}{\sum_{\omega' \in \Omega} \exp\big(\sum_{i=1}^{d} l_i(\varphi(\theta))v_i(\omega')\big)}, \qquad \omega \in \Omega.$$

With

$$D_{\mathcal{E}}\big(\varphi(\theta)\big) = \sum_{i=1}^{n} \theta_i \ln \frac{\theta_i}{\pi_{\mathcal{E}}(\varphi(\theta))(\omega_i)} + \left(1 - \sum_{i=1}^{n} \theta_i\right) \ln \frac{\big(1 - \sum_{i=1}^{n} \theta_i\big)}{\pi_{\mathcal{E}}(\varphi(\theta))(\omega_{n+1})},$$

elementary calculations lead to

$$\begin{aligned} \partial_i D_{\mathcal{E}}(p) &= \frac{\partial (D_{\mathcal{E}} \circ \varphi)}{\partial \theta_i}\big(\varphi^{-1}(p)\big) \\ &= \ln \frac{p(\omega_i)}{\pi_{\mathcal{E}}(p)(\omega_i)} - \ln \frac{p(\omega_{n+1})}{\pi_{\mathcal{E}}(p)(\omega_{n+1})} \\ &\quad + \sum_{j=1}^{d} \frac{\partial (l_j \circ \varphi)}{\partial \theta_i}(\theta)\big(\mathbb{E}_{\pi_{\mathcal{E}}(p)}(v_j) - \mathbb{E}_p(v_j)\big). \end{aligned} \tag{4.11}$$

It is well known that the equation

$$\mathbb{E}_{\pi_{\mathcal{E}}(p)}(x) = \mathbb{E}_p(x) \tag{4.12}$$

holds for all $x \in \mathrm{V}$. Thus the term (4.11) vanishes, and with the gradient formula we get

$$\begin{aligned} \operatorname{grad} D_{\mathcal{E}}(p) &= \sum_{i,j=1}^{n} g^{ij}(p)\partial_i D_{\mathcal{E}}(p)\partial_j(p) \\ &= \sum_{i,j=1}^{n} p(\omega_i)\big(\delta_{ij} - p(\omega_j)\big)\left(\ln \frac{p(\omega_i)}{\pi_{\mathcal{E}}(p)(\omega_i)} - \ln \frac{p(\omega_{n+1})}{\pi_{\mathcal{E}}(p)(\omega_{n+1})}\right)\partial_j(p) \\ &= \sum_{i=1}^{n} p(\omega_i)\left(\ln \frac{p(\omega_i)}{\pi_{\mathcal{E}}(p)(\omega_i)} - D_{\mathcal{E}}(p)\right)\partial_i(p) \\ &= \sum_{\omega \in \Sigma} p(\omega)\left(\ln \frac{p(\omega)}{\pi_{\mathcal{E}}(p)(\omega)} - D_{\mathcal{E}}(p)\right)e_\omega. \end{aligned}$$

□

PROOF OF PROPOSITION 3.2. With $\operatorname{aff} \mathrm{C}$ we denote the usual affine hull of a set $\mathrm{C} \subset \mathbb{R}^{\Omega}$. We set $\mathcal{P} := \mathcal{P}(\operatorname{supp} p)$ and $\mathrm{F} := \operatorname{aff} \pi_{\mathcal{E}}^{-1}(\{\pi_{\mathcal{E}}(p)\})$ and define the (-1)-convex set

$$\mathrm{S} := \mathrm{F} \cap \mathcal{P} \subset \operatorname{dom} \mathcal{E}.$$

With $D_{\mathcal{E}}$, the restriction of $D_{\mathcal{E}}$ to S attains a locally maximal value in p. Because of the strict convexity of this restriction, p is an extreme point of S. Furthermore, S is a relatively open set (i.e., open in $\operatorname{aff} S$). This is only possible in the case in which S consists of exactly one point: $S = \{p\}$. With this, one also has

$$F \cap \operatorname{aff} \mathcal{P} = \{p\}. \tag{4.13}$$

Finally, we apply the dimension formula:

$$\begin{aligned} |\Omega| - 1 &= \dim \mathcal{P}(\Omega) = \dim \operatorname{aff} \mathcal{P}(\Omega) \geq \dim(F \cup \operatorname{aff} \mathcal{P}) \\ &= \dim F + \dim \operatorname{aff} \mathcal{P} - \underbrace{\dim(F \cap \operatorname{aff} \mathcal{P})}_{=0, \text{ with (4.13)}} \\ &= (|\Omega| - 1 - \dim \mathcal{E}) + |\operatorname{supp} p| - 1. \end{aligned}$$

□

PROOF OF THEOREM 3.5. Let $\mathcal{E}$ be generated by p_0 and $v_1, \ldots, v_d$ and $p \in \operatorname{dom} \mathcal{E}$ a locally maximal point with respect to $D_{\mathcal{E}}$. From Proposition 3.2 we know $|\operatorname{supp} p| \leq d + 1$. Now we choose an injective map $\varphi = (\varphi_1, \ldots, \varphi_{d+1})\colon \Omega \to \mathbb{R}^{d+1}$ such that the points $\varphi(\omega)$, $\omega \in \Omega$, are in general position; that is, d' elements of $\varphi(\Omega)$ with $d' \leq d + 2$ are affinely independent. This guarantees the existence of real numbers $a_1, \ldots, a_{d+1}, b$ such that

$$\left\{ \omega \in \Omega : \sum_{i=1}^{d+1} a_i \varphi_i(\omega) = b \right\} = \operatorname{supp} p \tag{4.14}$$

holds.

Define $\mathcal{E}^*$ to be generated by p_0 and

$$v_1, \ldots, v_d, \quad \varphi_1, \ldots, \varphi_{d+1}, \quad \varphi_i \varphi_j, 1 \leq i \leq j \leq d + 1.$$

We have

$$\dim \mathcal{E}^* \leq d + (d+1) + \frac{(d+1)^2 + (d+1)}{2} = \frac{1}{2}(d^2 + 7d + 4).$$

Finally, we show that there is a sequence (p_n) in $\mathcal{E}^*$ that converges to p. With a sequence $\beta_n \uparrow \infty$ and real numbers $a_1, \ldots, a_{d+1}, b$ satisfying (4.14), we set

$$x := \ln \frac{\pi_{\mathcal{E}}(p)}{p_0}, \qquad r_i^{(n)} := 2\beta_n b a_i, \qquad s_{ij}^{(n)} := -\beta_n a_i a_j,$$

$$x_n := x + \sum_{i=1}^{d+1} r_i^{(n)} \varphi_i + \sum_{i,j=1}^{d+1} s_{ij}^{(n)} \varphi_i \varphi_j - \beta_n b^2 = x - \beta_n \left(\sum_{i=1}^{d+1} a_i \varphi_i - b \right)^2$$

and

$$p_n := \frac{p_0 \exp x_n}{\sum_{\omega' \in \Omega} p_0(\omega') \exp x_n(\omega')} \in \mathcal{E}^*.$$

With this, we have

$$\lim_{n\to\infty} p_n(\omega) = \frac{\pi_{\mathcal{E}}(p)(\omega)}{\pi_{\mathcal{E}}(p)(\operatorname{supp} p)}\mathbb{I}_{\operatorname{supp} p}(\omega) = p(\omega).$$

Here, the last equality follows from (3.3). □

PROOF OF LEMMA 3.7. With (4.12), we have

$$p(\operatorname{supp} p \cap A) = p(A) = \pi_{\mathcal{E}_{\mathscr{A}}}(p)(A) > 0$$

and therefore $\operatorname{supp} p \cap A \neq \varnothing$ for every nonempty set $A \subset \Omega$ with $A \in \mathscr{A}$ or $\Omega \setminus A \in \mathscr{A}$. □

PROOF OF PROPOSITION 3.9. The first statement about the existence and uniqueness of maximal solutions follows from the regularity of the vector field (see [15], page 159). To prove the second statement, assume $\sup I < \infty$. Then $p := \lim_{\tau\to\sup I} \gamma(\tau)$ is a projectable element of the boundary $\bar{\mathcal{P}}(\Omega) \setminus \mathcal{P}(\Omega)$ (see [15], page 171). Let U be an open neighborhood of p in $\mathbb{R}^\Omega$ as in Lemma 4.1. Because of the continuous differentiability of the l_i, there exists an open ball $B \subset \mathbb{R}^\Omega$ with $p \in B \subset U$ such that the function $B \to \mathbb{R}$, $x \mapsto \max_{1\le i\le d} |l_i(x)|$, is bounded from above. Now consider the exponential map $\exp_c\colon \mathrm{T}(\Omega) \to \mathcal{P}(\Omega)$ in the centroid c with respect to the $(+1)$-geodesics:

$$x \mapsto \sum_{\omega\in\Omega} \frac{\exp x(\omega)}{\sum_{\omega'\in\Omega} \exp x(\omega')} e_\omega.$$

The preimage $V := \exp_c^{-1}(B \cap \mathcal{P}(\Omega))$ is an open and unbounded set in $\mathrm{T}(\Omega)$ and we get

$$\|(\operatorname{grad} D_{\mathcal{E}} \circ \exp_c)(x)\| \le \|x\| + \text{constant}$$

for all $x \in V$; that is, the composed vector field is linearly bounded on V. Therefore the solutions can be extended to all positive times. □

PROOF OF THEOREM 3.12. We know that there exists a solution $\gamma = (\gamma_\omega)_{\omega\in\Omega}\colon I \to \mathcal{P}(\Omega)$ for $\Psi_{\mathcal{E}}$ with $\sup I = \infty$ and $\lim_{t\to\infty} \gamma(t) = p$ (Proposition 3.9). As a continuous function on a compact set, $D_{\mathcal{E}}$ is bounded from above by a number $C < \infty$ and we have

$$\begin{aligned}\int_0^t \left\|\frac{d\gamma}{dt}(\tau)\right\|^2 d\tau &\le \int_0^t \left\|\frac{d\gamma}{dt}(\tau)\right\|^2_{\gamma(\tau)} d\tau \\ &= \int_0^t \left\langle \frac{d\gamma}{dt}(\tau), \frac{d\gamma}{dt}(\tau)\right\rangle_{\gamma(\tau)} d\tau \\ &= \int_0^t \left\langle \operatorname{grad} D_{\mathcal{E}}(\gamma(\tau)), \frac{d\gamma}{dt}(\tau)\right\rangle_{\gamma(\tau)} d\tau\end{aligned}$$

$$= D_{\mathcal{E}}\big(\gamma(t)\big) - D_{\mathcal{E}}\big(\gamma(0)\big)$$
$$\leq 2C.$$

This implies

$$\lim_{t\to\infty} \int_0^t \left\| \frac{d\gamma}{dt}(\tau) \right\|^2 d\tau < \infty,$$

and the sets

$$A(n,T) := \left\{ \tau > T : \left\| \frac{d\gamma}{dt}(\tau) \right\|^2 \leq \frac{1}{n} \right\}, \qquad n = 1, 2, \ldots, T \in I,$$

are nonempty. We choose a sequence $(\tau_n)_{n\in\mathbb{N}}$ of real numbers with

$$\tau_0 = 0 \quad \text{and} \quad \tau_n \in A\big(n, \max\{n, \tau_{n-1}\}\big), \qquad n = 1, 2, \ldots .$$

Such a sequence obviously has the properties $\tau_n \uparrow \infty$ and $\lim_{n\to\infty} \|(d\gamma/dt)(\tau_n)\| = 0$. Thus

$$\begin{aligned} 0 &= \lim_{n\to\infty} \frac{d\gamma_\omega}{dt}(\tau_n) \\ &= \lim_{n\to\infty} \gamma_\omega(\tau_n) \left(\ln \frac{\gamma_\omega(\tau_n)}{\pi_{\mathcal{E}}(\gamma(\tau_n))(\omega)} - D_{\mathcal{E}}\big(\gamma(\tau_n)\big) \right) \\ &= p(\omega) \left(\ln \frac{p(\omega)}{\pi_{\mathcal{E}}(p)(\omega)} - D_{\mathcal{E}}(p) \right). \end{aligned}$$

This proves $p(\omega) = e^{D_{\mathcal{E}}(p)} \pi_{\mathcal{E}}(p)(\omega)$ for all $\omega \in \operatorname{supp} p$.

To complete the proof, according to Proposition 3.2 it is sufficient to show that $D_{\mathcal{E}}$ attains a locally maximal value in the stable limit point p: The stability assumption guarantees the existence of a neighborhood U of p in $\bar{\mathcal{P}}(\Omega)$ such that for every initial point $p_0 \in U \cap \mathcal{P}(\Omega)$ there is a solution for the gradient field $\Psi_{\mathcal{E}}$ with $\lim_{t\to\infty} \gamma(t) = p$. Of course, we have $D_{\mathcal{E}}(p_0) \leq D_{\mathcal{E}}(p)$. The continuity of $D_{\mathcal{E}}$ implies that this inequality also holds for an arbitrary $p_0 \in U = (U \setminus \mathcal{P}(\Omega)) \uplus (U \cap \mathcal{P}(\Omega))$. □

Acknowledgments. I am grateful to Professor Jürgen Jost for his general support and to Ulrich Steinmetz for the implementation of the presented approach into computer programs (Main Example, Part 3).

REFERENCES

[1] AMARI, S.-I. (1985). *Differential-Geometric Methods in Statistics*. *Lecture Notes in Statist.* **28**. Springer, Berlin.

[2] AMARI, S.-I. (1997). Information geometry. *Contemp. Math.* **203** 81–95.

[3] AMARI, S.-I. (2001). Information geometry on hierarchy of probability distributions. *IEEE Trans. Inform. Theory* **47** 1701–1711.
[4] AMARI, S.-I. and NAGAOKA, H. (2000). *Methods of Information Geometry*. *Math. Monogr.* **191**. Oxford Univ. Press.
[5] AMARI, S.-I., BARNDORFF–NIELSEN, O. E., KASS, R. E., LAURITZEN, S. L. and RAO, C. R. (1987). *Differential Geometry in Statistical Inference*. IMS, Hayward, CA.
[6] AY, N. (2000). Aspekte einer Theorie pragmatischer Informationsstrukturierung. Ph.D. dissertation, Univ. Leipzig.
[7] BOOTHBY, W. M. (1975). An Introduction to Differentiable Manifolds and Riemannian Geometry. *Pure Appl. Math.* **63**. Academic Press, New York.
[8] BRONDSTED, A. (1983). *An Introduction to Convex Polytopes*. Springer, New York.
[9] COVER, T. M. and THOMAS, J. A. (1991). *Elements of Information Theory*. Wiley-Interscience, New York.
[10] CSISZÁR, I. (1967). On topological properties of f-divergence. *Studia Sci. Math. Hungar.* **2** 329–339.
[11] CSISZÁR, I. (1975). I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.* **3** 146–158.
[12] DECO, G. and OBRADOVIC, D. (1996). *An Information-Theoretic Approach to Neural Computing*. *Perspectives in Neural Computing*. Springer, New York.
[13] FUJIWARA, A. and AMARI, S.-I. (1995). Gradient systems in view of information geometry. *Phys. D* **80** 317–327.
[14] GZYL, H. (1995). *The Method of Maximum Entropy. Ser. Adv. Math. Appl. Sci.* **29**. World Scientific, Singapore.
[15] HIRSCH, M. and SMALE, S. (1974). *Differential Equations, Dynamical Systems, and Linear Algebra*. Academic Press, New York.
[16] INGARDEN, R. S., KOSSAKOWSKI A. and OHYA M. (1997). *Information Dynamics and Open Systems, Classical and Quantum Approach*. Kluwer, Dordrecht.
[17] JAYNES, E. T. (1957). Information theory and statistical mechanics. *Phys. Rev.* **106**.
[18] KULLBACK, S. (1968). *Information Theory and Statistics*. Dover, Mineola, NY.
[19] KULLBACK, S. and LEIBLER, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.* **22** 79–86.
[20] LINSKER, R. (1988). Self-organization in a perceptual network. *Computer* **21** 105–117.
[21] MARTIGNON, L., VON HASSELN, H., GRÜN, S., AERTSEN, A. and PALM, G. (1995). Detecting higher-order interactions among the spiking events in a group of neurons. *Biol. Cybernet.* **73** 69–81.
[22] MURRAY, M. K. and RICE, J. W. (1994). *Differential Geometry and Statistics.* Chapman and Hall, London.
[23] NAGAOKA, H. and AMARI, S. (1982). Differential geometry of smooth families of probability distributions. AETR 82-7, Univ. Tokyo.
[24] NAKAMURA, Y. (1993). Completely integrable gradient systems on the manifolds of gaussian and multinomial distributions. *Japan J. Indust. Appl. Math.* **10** 179–189.
[25] RAO, C. R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* **37** 81–91.
[26] ROCKAFELLAR, R. T. and WETS, J. B. R. (1998). *Variational Analysis.* Springer, New York.
[27] ROMAN, S. (1992). *Coding and Information Theory*. Springer, New York.
[28] SHANNON, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.* **27** 379–423, 623–656.
[29] VAPNIK, V. (1998). *Statistical Learning Theory. Adaptive and Learning Systems for Signal Processing, Communications, and Control*. Wiley, New York.

[30] VAPNIK, V. and CHERVONENKIS, A. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.* **16** 264–280.
[31] WEBSTER, R. (1994). *Convexity*. Oxford Univ. Press.

MAX-PLANCK-INSTITUTE FOR MATHEMATICS
IN THE SCIENCES
INSELSTRASSE 22-26
04103 LEIPZIG
GERMANY
E-MAIL: nay@mis.mpg.de