Max-Planck-Institut
für Mathematik
in den Naturwissenschaften
Leipzig

Locality of global stochastic interaction in directed acyclic networks

(revised version: March 2002)

by

Nihat Ay

Preprint no.: 54 2001

Locality of Global Stochastic Interaction in Directed Acyclic Networks

Nihat Ay

Max-Planck-Institute for Mathematics in the Sciences

Inselstr. 22-24

04103 Leipzig, Germany

E-mail: nay@mis.mpg.de

26th March 2002

Abstract

The hypothesis of invariant maximization of interaction (IMI) is formulated within the setting of random fields. According to this hypothesis, learning processes maximize the stochastic interaction of the neurons subject to constraints. We consider the extrinsic constraint in terms of a fixed input distribution on the periphery of the network. Our main intrinsic constraint is given by a directed acyclic network structure. First mathematical results about the strong relation of the local information flow and the global interaction are stated in order to investigate the possibility of controlling IMI optimization in a completely local way. Furthermore, we discuss some relations of this approach to the optimization according to Linsker's Infomax principle.

Key words and phrases. Infomax principle, stochastic interaction, directed acyclic networks, information geometry, random fields.

1 Introduction

The present paper is based on the hypothesis that neural networks are realizations of complex systems in the sense that their elementary units (neurons) are strongly interacting with each other, subject to constraints. The following challenging statement by Chaitin encourages to speculate on the fundamental subject of complexity from a very general point of view (2001, p. 97):

"Many years ago I used mutual information in an attempt to formulate a general mathematical definition of life, to distinguish organized living matter from ordinary matter. The idea was that the parts of a living organism are highly correlated, highly interdependent, and have high mutual information. After all, primordial quality of living beings is their tremendous complexity and interdependentness. I never got very far with this theory. Can you do better?"

Within the field of neural networks several complexity measures related to the one by Chaitin (1979) have been proposed. Tononi, Sporns, and Edelman (1994) introduced a measure for brain complexity that is intended to describe segregation and integration in the brain in terms of mutual information with respect to all bipartitions. Jost (2000/2001) generalized it to the case of arbitrary partitions, where higher order interactions among more than two subsystems are included. In the present approach we formalize the preceding hypothesis by using a notion of complexity based on the stochastic interaction among the elementary subsystems,

namely the neurons. It quantifies the stochastic interaction in a neural system with respect to a joint probability distribution p in terms of the Kullback-Leibler "distance" (Kullback & Leibler, 1951) of p from the set of all factorizable distributions. Amari (2001) discussed this measure from the information-geometric point of view. He used the "Pythagorean Theorem" (Amari & Nagaoka, 2000, p. 62f) to present a hierarchical decomposition of global stochastic interaction in lower-order interactions. In our previous work (Ay, 2001a), we studied probability distributions that maximize the global interaction without any constraints. The present paper continues this work and applies it to the field of neural networks.

Experimental evidence for the optimization of information flow in neural systems is mainly provided on the local level in terms of the mutual information between a neuron and its neighbours. Early experimental results by Laughlin (1981) concerning the response characteristics of large monopolar cells (LMC) in the visual system of the fly suggest that the local information transmission is maximized in a neural system. A detailed discussion on this topic is presented in the book by Rieke, Warland, Ruytervan Steveninck, and Bialek (1998). Linsker (1986, 1988) pointed out that local optimization can be achieved by Hebb-like learning. Hye convincingly demonstrated by computer simulations that this kind of local learning leads to the emergence of receptive fields in feed-forward networks which are surprisingly similar to those observed in the visual pathway of the cat and the monkey (Hubel & Wiesel, 1962, 1968). Motivated by the simulation

results, he formulated the *principle of maximum information preservation* (*Infomax*) for unsupervised or self-organized learning in the field of neural networks. In the present paper, we investigate the possibility of describing the optimization of local information transmission in directed acyclic networks in terms of the global interaction complexity.[1] Such a local-global compatibility would provide an attractive way to formulate an invariant first principle for information processing in neural systems based on our main hypothesis of strong global interaction.

The paper is organized as follows. Section 2 contains a brief introduction into the setting of random fields on directed acyclic graphs. Section 3.1 formalizes the hypothesis of strong interaction and applies it to the situation of our previous work (Ay, 2001a). Section 3.2 discusses the relation of our hypothesis to the Linsker Infomax. Section 3.3 presents some locality properties of global interaction in directed acyclic networks. Section 4.1 continues the discussion of Linsker's work in view of the results of our paper. We conclude the paper with some problems and comments in Section 4.2. The Appendix contains the proofs of the mathematical statements of the paper.

[1] A generalization of this approach to recurrent networks is in progress.

2 Preliminaries:

Random Fields on Directed Acyclic Graphs

We use the framework of random fields on directed acyclic networks which is recalled in what follows (see Lauritzen, 1996; Cowell et al, 1999).

(i) Random fields: The set of all probability distributions on a non-empty finite set Ω is denoted by $\bar{\mathcal{P}}(\Omega) \subset \mathbb{R}^{\Omega}$. The *support* of a probability distribution $p \in \bar{\mathcal{P}}(\Omega)$ is defined as $\operatorname{supp} p := \{\omega \in \Omega \,:\, p(\omega) > 0\}$. The strictly positive distributions $\mathcal{P}(\Omega)$ have the maximal support Ω.

Let V be a finite set of *neurons*. To each neuron $v \in V$ we assign a finite set Ω_v of *states*. For a subset $A \subset V$, the set of all configurations on A is given by the product $\Omega_A := \times_{v \in A} \Omega_v$. If $A = \emptyset$, then Ω_A consists of exactly one element, namely the empty configuration ϵ. The elements of $\bar{\mathcal{P}}(\Omega_A)$ are the *random fields* on A. One has the natural restriction $X_A : \Omega_V \to \Omega_A$, $(\omega_v)_{v \in V} \mapsto \omega_A := (\omega_v)_{v \in A}$, which induces the projection $\bar{\mathcal{P}}(\Omega_V) \to \bar{\mathcal{P}}(\Omega_A)$, $p \mapsto p_A$, where p_A denotes the image measure of p under the variable X_A.

(ii) Graphical structure: In the present paper we investigate random fields that are compatible with directed acyclic graphs. Such graphs are considered to be a general model for feed-forward neural networks. The set E of edges (*synapses*) is a subset of the set $V \times V$ of ordered pairs of neurons. The graph $N := (V, E)$ is called *directed and acyclic* iff for all sequences $v_0, \ldots, v_n \in V$ with

$(v_{k-1}, v_k) \in E$, $k = 1, \ldots, n$, the first neuron v_0 is different from the last neuron v_n. With each $v \in V$, we associate the set

$$\mathrm{pa}(v) \;:=\; \{w \in V \;:\; (w, v) \in E\}$$

of *parents* of v. We define the *periphery* of the graph N as

$$\mathrm{per}(N) \;:=\; \{v \in V \;:\; \mathrm{pa}(v) = \emptyset\}.$$

It is non-empty if N is directed and acyclic. The complement $V \setminus \mathrm{per}(N)$ of the periphery is denoted by $\mathrm{int}(N)$.

(iii) Combination of (i) and (ii): Now we introduce random fields on V that are compatible with the structure of a given directed and acyclic graph. First of all, we assume that there is a probability distribution $p^\partial \in \bar{\mathcal{P}}\big(\Omega_{\mathrm{per}(N)}\big)$ on the network's periphery that is independent from intrinsic properties of the system. Furthermore we assume that for each neuron $v \in \mathrm{int}(N)$ there is a local kernel function

$$K_v : \Omega_{\mathrm{pa}(v)} \times \Omega_v \to [0, 1], \qquad (\omega, \omega') \mapsto K_v(\omega' \,|\, \omega),$$

with

$$\sum_{\omega' \in \Omega_v} K_v(\omega' \,|\, \omega) \;=\; 1, \qquad \text{for all } \omega \in \Omega_{\mathrm{pa}(v)}.$$

The distribution p^∂ and the family $(K_v)_{v \in \mathrm{int}(N)}$ of kernels define the composed distribution

$$p(\omega) \;:=\; p^\partial(\omega_{\mathrm{per}(N)}) \prod_{v \in \mathrm{int}(N)} K_v(\omega_v \,|\, \omega_{\mathrm{pa}(v)}), \qquad \omega \in \Omega_V. \tag{1}$$

We denote it by $p^\partial \otimes \left(\otimes_{v\in \mathrm{int}(N)} K_v\right)$ and say that it is (N, p^∂)-*adapted*. Let $\bar{\mathcal{K}}_v$ be the set of local kernel functions for the neuron v, and let $\mathcal{K}_v$ be its subset of strictly positive kernel functions. Then we have the composition map

$$\bar{\mathcal{P}}(\Omega_\partial) \times \prod_{v\in \mathrm{int}(N)} \bar{\mathcal{K}}_v \;\hookrightarrow\; \bar{\mathcal{P}}(\Omega_V), \qquad (p^\partial; K_v, v \in \mathrm{int}(N)) \;\mapsto\; p^\partial \otimes \left(\otimes_{v\in \mathrm{int}(N)} K_v\right).$$

The image of this map is denoted by $\mathcal{C}(N, p^\partial)$. The elements of $\mathcal{C}(N, p^\partial)$ are exactly the (N, p^∂)-adapted probability distributions.

Remark 2.1. In the present paper we use the term *local* with two different meanings, which can be clearly distinguished from the context. The first meaning refers to the definition of a local maximizer of a function. The second way to use this term describes what is usually meant by *local information flow* or *local learning* in a neural network, and therefore depends on the underlying network structure.

3 Invariant Maximization of Interaction

3.1 The Formalization of the Hypothesis

As stated in the introduction, the present paper is based on the hypothesis that neural networks are realizations of systems with strongly interacting units. Now we formalize this approach using well-known information-theoretic quantities.

For two disjoint subsystems $A, B \subset V$, and a probability distribution $p \in \bar{\mathcal{P}}(\Omega_V)$, we define the *entropy* on A by

$$H_p(A) \;:=\; -\sum_{\omega \in \Omega_A} p(X_A = \omega)\, \ln p(X_A = \omega)$$

and the *conditional entropy* on B given A, by

$$H_p(B \,|\, A) \;:=\; -\sum_{\omega \in \Omega_A,\, \omega' \in \Omega_B} p(X_A = \omega, X_B = \omega')\, \ln p(X_B = \omega' \,|\, X_A = \omega).$$

The *mutual information* or *transinformation* of A and B is defined as

$$I_p(A;B) \;:=\; H_p(A) + H_p(B) - H_p(A \uplus B).$$

The (*stochastic*) *interaction* of the neurons in A is defined in a similar way:

$$I_p^A \;:=\; \sum_{v \in A} H_p(v) - H_p(A).$$

We have

$$I_p^{A \uplus B} \;=\; I_p^A + I_p^B + I_p(A;B). \tag{2}$$

Without specifying the probability distribution p, we use the same notation for the corresponding functions $\bar{\mathcal{P}}(\Omega_V) \to \mathbb{R}_+$. For example, $H(A)$ denotes the function $p \mapsto H_p(A)$.

In these definitions, we assume that learning in neural systems is driven by mechanisms that maximize the stochastic interaction I^V subject to constraints. The formulation of this hypothesis does not require any specification of the underlying constraints. For this reason, we call it *the hypothesis of **I**nvariant **M**aximization of **I**nteraction (IMI).*

The completely unconstrained maximization of I^V leads to a strong reduction of the number of possible configurations in the system. Ay (2001a) studied this reduction and proved the following proposition.

PROPOSITION 3.1. *Let p be a local maximizer of I^V. Then the following bound on the support of p holds:*

$$|\operatorname{supp} p| \;\leq\; 1 + \sum_{v \in V} \big(|\Omega_v| - 1\big). \tag{3}$$

Thus, for binary neurons (i.e., $|\Omega_v| = 2$ for all $v \in V$) the number of configurations is reduced from $2^{|V|}$ to at most $|V|+1$. This is illustrated by the following examples obtained from computer simulations.

Example 3.2. Each Venn diagram in the Figures 1 and 2 represents a probability distribution on the set of all atoms that are generated by the corresponding events. The grey value of an atom is proportional to the probability of the atom. "White" corresponds to the maximal probability in a given Venn diagram. The diagrams on the left-hand side are initial distributions which induce those on the right-hand side by unconstrained natural gradient flow (Amari, 1998) that optimizes the stochatic interaction.

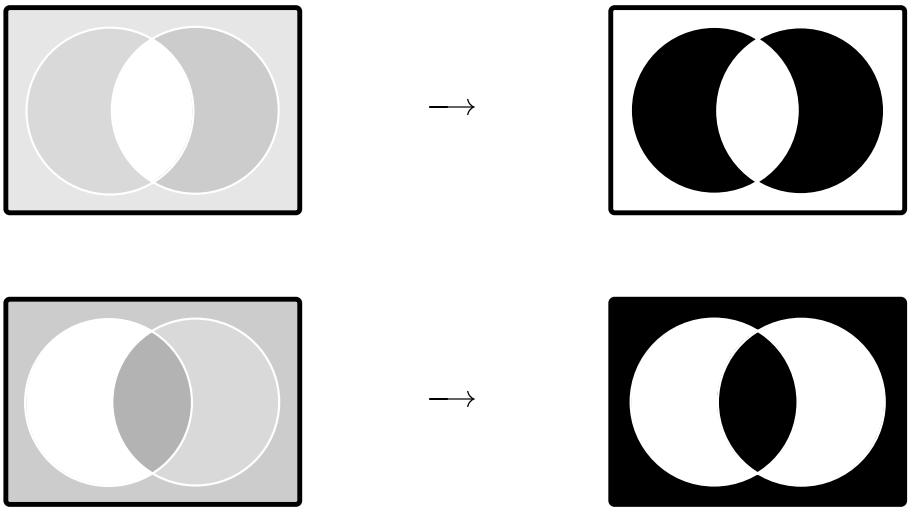

Fig. 1.

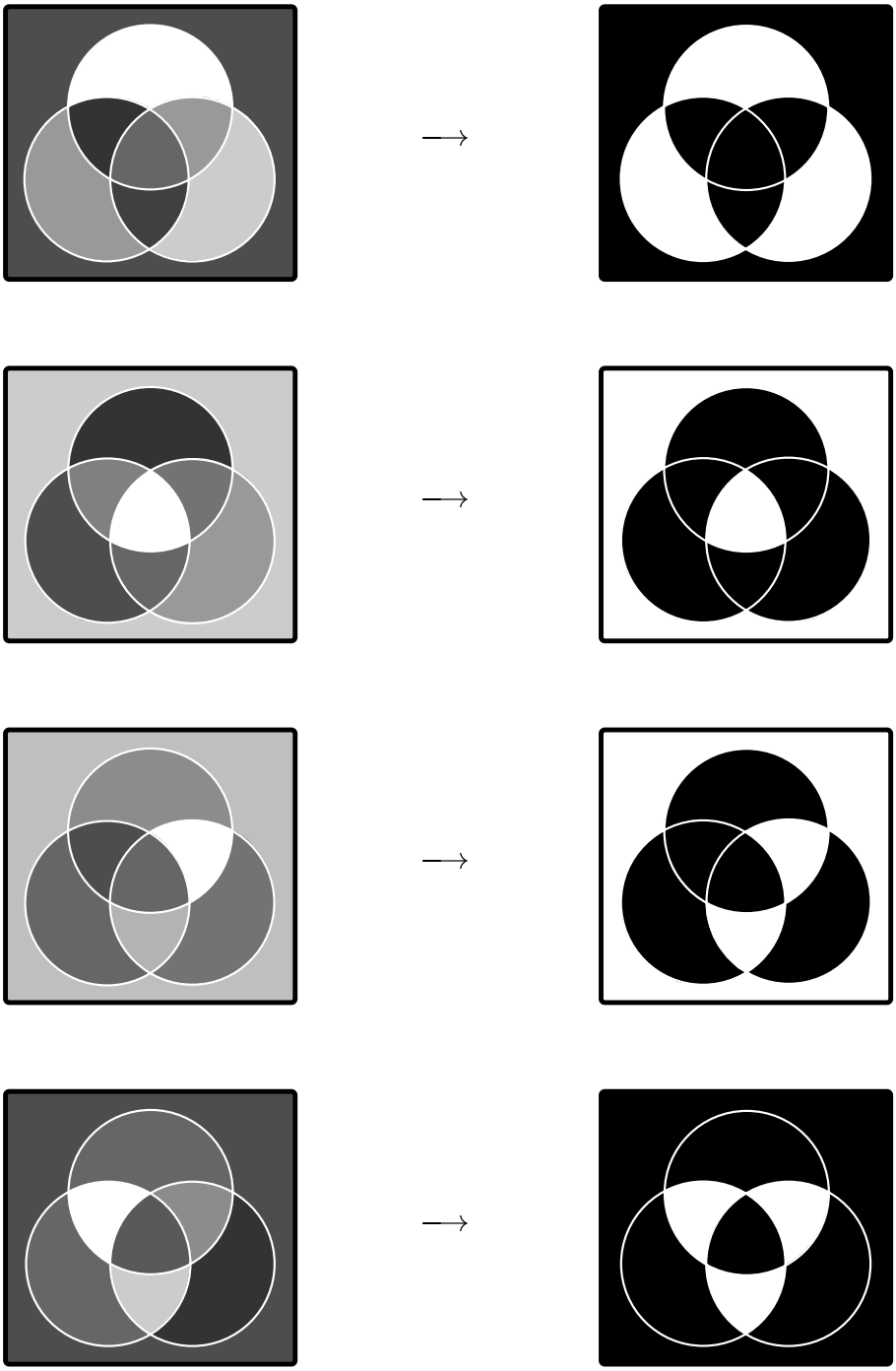

FIG. 2.

In this paper we investigate the maximization of interaction with respect to *extrinsic* and *intrinsic* constraints. The extrinsic constraint is modelled by a fixed input distribution on the periphery of the system. We get a restriction of the optimization to a convex subset of the set of probability distributions which is discussed in Section 3.2. The specification of the internal structure of the system restricts the optimization to a more complicated manifold which describes the intrinsic constraints. We have to consider two aspects:

1. Specification of the network structure in terms of a graph;
2. Specification of the system parametrization that is compatible with the network structure.

Intrinsic constraints are investigated in Section 3.3.

3.2 The Extrinsic Constraint and Linsker's Infomax

Let ∂ be a subset of the set V of neurons, called *periphery* of V, and p^∂ be a probability distribution on Ω_∂. By $\mathcal{C}(p^\partial)$ we denote the convex set of probability distributions q on Ω_V that satisfy

$$q(X_\partial = \omega) \;=\; p^\partial(X_\partial = \omega) \quad \text{for all } \omega \in \Omega_\partial.$$

Using equation (2), we can split the global stochastic interaction in the following way:

$$I^V \;=\; I^\partial \;+\; I(\partial; V \setminus \partial) \;+\; I^{V\setminus\partial}. \tag{4}$$

Note that the first term on the right-hand side of equation (4) depends only on the peripheral distribution p^∂. Therefore, it is constant on the set $\mathcal{C}(p^\partial)$ of distributions with ∂-marginal p^∂.

$$I_p^V \;=\; I_p(\partial; V \setminus \partial) \;+\; I_p^{V\setminus\partial} \;+\; const, \qquad\qquad p \in \mathcal{C}(p^\partial). \tag{5}$$

Thus, the $\mathcal{C}(p^\partial)$-constrained optimization of interaction can be thought of as a combination of two optimizations. The first one corresponds to

$$I_p(\partial; V \setminus \partial) \;=\; H_p(V \setminus \partial) - H_p(V \setminus \partial \,|\, \partial) \tag{6}$$

and is closely related to Linsker's Infomax principle (Linsker 1988), which states that learning processes in layered networks maximize the transinformation between the input and the output (we discuss Linsker's work in Section 4.1). In many models, this principle leads to a strong decorrelation of the output neurons (Bell & Sejnowski, 1995; Obradovic & Deco, 1998). Therefore, it is considered to be consistent with the concept of redundancy reduction by Attneave (1954) and Barlow (1989). On the other hand, the internal integration of information is based on functional dependences and correlations, so that the optimization according to the Infomax priciple should be controlled by the maximization of the redundancy of the output neurons (this has also been proposed by Barlow,

2001). Combining the Infomax term $I_p(\partial; V \setminus \partial)$ with the redundancy term $I_p^{V\setminus\partial}$, we end up with the optimization of the right-hand side of (5), which is equivalent on $\mathcal{C}(p^\partial)$ to the maximization of interaction.

So far we have compared the IMI hypothesis with Linsker's Infomax principle from the conceptual point of view. There is another way to compare these two approaches. In order to do this, we use (5) to rewrite the global interaction on $\mathcal{C}(p^\partial)$:

$$\begin{aligned} I_p^V &= \Big(H_p(V \setminus \partial) - H_p(V \setminus \partial \,|\, \partial)\Big) \;+\; \left(\sum_{v\in V\setminus\partial} H_p(v) - H_p(V \setminus \partial)\right) \;+\; const \\ &= \sum_{v\in V\setminus\partial} H_p(v) \;-\; H_p(V \setminus \partial \,|\, \partial) \;+\; const. \end{aligned} \tag{7}$$

Comparing (7) with (6), we observe that the main difference between the global interaction and the transinformation on $\mathcal{C}(p^\partial)$ is that the output entropy $H_p(V\setminus\partial)$ in (6) is replaced by the sum $\sum_{v\in V\setminus\partial} H_p(v)$ of the individual "local" output entropies. Under certain assumptions, the conditional entropy $H_p(V \setminus \partial \,|\, \partial)$ in (7) also localizes and provides a way to define local learning rules for the optimization of interaction. We study this nice locality property in Section 3.3. Furthermore, we observe that both the Infomax and the IMI optimizations have the tendency to produce input-output relations with low conditional entropy. The following generalization of Proposition 3.1 gives us an estimate of the conditional entropy with respect to a distribution that maximizes the interaction on $\mathcal{C}(p^\partial)$.

PROPOSITION 3.3. *Let ∂ be a subset of V, and let p^∂ be a probability distribution on Ω_∂. If $p \in \bar{\mathcal{P}}(\Omega_V)$ is a local maximizer of I^V in the convex set $\mathcal{C}(p^\partial)$ of distributions with ∂-marginal p^∂, then*

$$0 \;\leq\; |\operatorname{supp} p| - |\operatorname{supp} p^\partial| \;\leq\; \sum_{v \in V \setminus \partial} \big(|\Omega_v| - 1\big). \tag{8}$$

In particular, for binary neurons, inequality (8) implies

$$0 \;\leq\; |\operatorname{supp} p| - |\operatorname{supp} p^\partial| \;\leq\; |V \setminus \partial|.$$

Note that we recover the estimate (3) if we set $\partial := \emptyset$ in (8).

COROLLARY 3.4. *In the situation of Proposition 3.3, for all $\omega \in \operatorname{supp} p^\partial$*

$$|\operatorname{supp} p(\cdot \,|\, X_\partial = \omega)| \;\leq\; 1 + \sum_{v \in V \setminus \partial} \big(|\Omega_v| - 1\big).$$

COROLLARY 3.5. *In the situation of Proposition 3.3, the conditional entropy of the internal state under the condition of the external state satisfies*

$$H_p(V \setminus \partial \,|\, \partial) \;\leq\; \ln \left(1 + \sum_{v \in V \setminus \partial} \big(|\Omega_v| - 1\big) \right).$$

From Corollaries 3.4 and 3.5 we observe that the optimization of interaction with a fixed input distribution leads to a nearly functional dependency between

the input and the output. Consider for example the binary case. Given an input configuration on the periphery ∂, there are in general $2^{|V\setminus\partial|}$ possible outputs. According to Corollary 3.4, this maximal number is reduced to at most $1+|V\setminus\partial|$ configurations if we have a distribution that maximizes the interaction of the neurons in the set $\mathcal{C}(p^\partial)$. Thus, we have an exponential reduction of the possible outputs. As stated in Corollary 3.5, the corresponding conditional entropy that quantifies the uncertainty about the output, given the input, is bounded from above by $\ln\big(1+|V\setminus\partial|\big)$.

3.3 The Intrinsic Constraints and the Locality of Learning

To motivate the main idea of this section, we consider the example where all neurons except one, $v_0 \in V$, are elements of the periphery: $\partial = V \setminus v_0$. In this case, formula (4) becomes

$$I^V \;=\; I^\partial \;+\; I(v_0; V\setminus v_0). \tag{9}$$

Furthermore, consider a peripheral distribution p^∂ and the directed graph $N = (V, E)$, where E denotes the set of edges that go from the periphery to the neuron v_0, that is $E = \partial \times \{v_0\}$. Then (9) implies that all (N, p^∂)-adapted probability distributions p (see Section 2 (iii)) satisfy

$$I_p^V \;=\; I_{p^\partial}^\partial \;+\; I_p\big(v_0; \mathrm{pa}(v_0)\big). \tag{10}$$

This formula states that the optimization of the global interaction I^V in the set of (N, p^∂)-adapted distributions is equivalent to the optimization of the transinformation between the neuron v_0 and its parents $\mathrm{pa}(v_0)$, and therefore also equivalent to the optimization according to Linsker's Infomax principle. In the field of neural networks the hypothesis that neurons optimize information about their local environment is the subject of many theoretical and experimental investigations (Laughlin, 1981; Rieke, Warland, Ruytervan Steveninck, and Bialek, 1998). On the other hand, it is important to translate this hypothesis of local information maximization into an optimization principle that corresponds to a globally defined complexity measure (Tononi, Sporns, and Edelman, 1994). In the present paper, we prove some locality properties of the IMI-optimization for directed acyclic networks. The general theory, which also describes recurrent networks and establishes the equivalence of the maximization of local interactions to the maximization of global complexity, will be presented in another paper.

Now we extend the representation (10) to more than one inernal neuron.

PROPOSITION 3.6. *Let $N = (V, E)$ be a directed and acyclic graph with periphery $\partial = \mathrm{per}(N)$, and let p^∂ be a probability distribution on Ω_∂. If a probability distribution p is (N, p^∂)-adapted in the sense of Section 2 (iii), then*

$$I_p^V \;=\; I_{p^\partial}^\partial \;+\; \sum_{v \in \mathrm{int}\,(N)} I_p\big(v; \mathrm{pa}(v)\big).$$

Thus, the global stochastic interaction in a directed acyclic network N can be expressed, up to a constant, as the sum of local interactions of the neurons with their neighbours. In particular, this is the case in a multi-layer network without any lateral connections within the individual layers. Therefore the optimization of the local interactions in a directed acyclic graph is sufficient for the optimization of the global interaction. This indicates the possibility of controlling learning processes in terms of local interactions. This local-global connection can be easily investigated in networks without any hidden part, which is the subject in what follows. Networks of this type are called *simple*. More precisely, a directed acyclic network $N = (V, E)$ is simple if $\mathrm{pa}(v) \subset \mathrm{per}(N)$ holds for all $v \in V$. A simple network consists of the two *layers* $\mathrm{per}(N)$ and $\mathrm{int}(N) = V \setminus \mathrm{per}(N)$ where all connections go from $\mathrm{per}(N)$ to $\mathrm{int}(N)$: $E \subset \mathrm{per}(N) \times \mathrm{int}(N)$. In particular, there are no *lateral* connections inside each layer. In simple networks, the following holds:

PROPOSITION 3.7. *Let $N = (V, E)$ be a simple network, and let p^∂ be a probability distribution on $\mathrm{per}(N)$, and let p be an (N, p^∂)-adapted distribution, that is $p \in \mathcal{C}(N, p^\partial)$. Then p is a local maximizer of I^V in $\mathcal{C}(N, p^\partial)$ if and only if p is a local maximizer of $I\big(v; \mathrm{pa}(v)\big)$ in $\mathcal{C}(N, p^\partial)$ for all $v \in \mathrm{int}(N)$. This statement remains true if "local maximizer" is replaced by "isolated local maximizer".*

Now we describe intrinsic constraints that are not only given by the network structure but also by the specification of a model that is compatible with this structure (see the comment on the specifications of intrinsic constraints at the end of Section 3.1). Therefore, we consider a simple network $N = (V, E)$ and choose an arbitrary numbering $\mathrm{int}(N) = \{v_1, \ldots, v_m\}$. Furthermore, we assume that there is a probability distribution p^∂ on the periphery $\partial = \mathrm{per}(N)$ of N and a family of parametrizations (embeddings) $k_r : \Theta_r \to \mathcal{K}_{v_r}$, where Θ_r is an open subset of $\mathbb{R}^{d_r}$. The *local parameters* $\theta^{(r,i)}$, $i = 1, \ldots, d_r$, determine the local kernel of the neuron v_r. For each $\theta^{(r)} \in \Theta_r$, the kernel function $k_r(\theta^{(r)})$ is written as $(\omega, \omega') \mapsto k_r\left(\omega' \,|\, \omega, \theta^{(r)}\right)$. Using p^∂ and this family of local parametrizations, we define the global parametrization φ that assigns to each $\theta = (\theta^{(1)}, \ldots, \theta^{(m)}) \in \Theta := \prod_{r=1}^m \Theta_r$ the composed probability distribution with

$$p(\omega \,|\, \theta) \;=\; p^\partial(\omega_\partial) \prod_{r=1}^{m} k_r\left(\omega_{v_r} \,|\, \omega_{\mathrm{pa}(v_r)}, \theta^{(r)}\right), \qquad \omega \in \Omega_V. \tag{11}$$

Variational *learning* in neural networks corresponds to an optimizing stochastic process in the parameter space that is usually derived from the gradient of

the corresponding utility function by adaptation to available information (non-anticipation, locality). In the present paper, we do not derive such learning processes. We investigate the local-global relation of stochastic interaction entirely in terms of the natural gradient (Amari, 1998) and in this context also talk about *learning*. The natural gradient of a function f on the image $\mathcal{N}$ of the parametrization φ is determined by the first fundamental form $G := \big(g_{(r,i)(s,j)}\big)_{(r,i)(s,j)}$ of the Fisher metric with

$$g_{(r,i)(s,j)}(\theta) \;:=\; \mathbb{E}_{\varphi(\theta)}\left(\frac{\partial \ln \varphi}{\partial \theta^{(r,i)}}(\theta)\,\frac{\partial \ln \varphi}{\partial \theta^{(s,j)}}(\theta)\right), \qquad \theta \in \Theta.$$

The local coordinates of the natural gradient of f are given by

$$\big(G^{-1}\,\nabla\,(f \circ \varphi)\big)(\theta), \qquad \theta \in \Theta,$$

where ∇ denotes the canonical gradient in $\bigoplus_{r=1}^{m} \mathbb{R}^{d_r} \cong \mathbb{R}^d$, $d = \sum_{r=1}^{m} d_r$.

Note that in simple networks the functions

$$I_\varphi\big(v_r; \mathrm{pa}(v_r)\big) : \theta \mapsto I_{\varphi(\theta)}\big(v_r; \mathrm{pa}(v_r)\big) \quad \text{and} \quad H_\varphi\big(\mathrm{pa}(v_r) \,|\, v_r\big) : \theta \mapsto H_{\varphi(\theta)}\big(\mathrm{pa}(v_r) \,|\, v_r\big)$$

do not depend on the non-local parameters $\theta^{(s,j)}$, $s \neq r$. The corresponding local functions $\Theta_r \to \mathbb{R}$ are denoted by $I^{loc}\big(v_r; \mathrm{pa}(v_r)\big)$ and $H^{loc}\big(\mathrm{pa}(v_r) \,|\, v_r\big)$.

THEOREM 3.8. *Let N be a simple network, and let φ be a parametrization of the form (11) with image $\mathcal{N}$. Then the coordinates of the natural gradient of I^V on $\mathcal{N}$ with respect to φ are split in a completely local way. More precisely, the following holds: For all r, the matrix $G_r := \big(g_{(r,i)(r,j)}\big)_{i,j}$, depends only on the local parameter vector $\theta^{(r)}$, and*

$$\begin{aligned}\big(G^{-1}\nabla(I^V\circ\varphi)\big)(\theta) &= \begin{pmatrix} \big(G_1^{-1}\ \nabla I^{loc}\big(v_1;\mathrm{pa}(v_1)\big)\big)\,(\theta^{(1)}) \\ \vdots \\ \big(G_m^{-1}\ \nabla I^{loc}\big(v_m;\mathrm{pa}(v_m)\big)\big)\,(\theta^{(m)}) \end{pmatrix} \\ &= -\begin{pmatrix} \big(G_1^{-1}\ \nabla H^{loc}\big(\mathrm{pa}(v_1)\,|\,v_1\big)\big)\,(\theta^{(1)}) \\ \vdots \\ \big(G_m^{-1}\ \nabla H^{loc}\big(\mathrm{pa}(v_m)\,|\,v_m\big)\big)\,(\theta^{(m)}) \end{pmatrix}.\end{aligned}$$

Informally, Theorem 3.8 can be stated as follows:

In simple networks, global learning according to the IMI hypothesis is equivalent to the local learning according to a family of local potential functions: Each neuron maximizes the flow of information from its parents.

According to Theorem 3.8, in a simple network it is sufficient to compute separately the gradients of the local information flows. The following example discusses a specific parametrization of a simple network with a single output neuron:

EXAMPLE 3.9. (SIMPLE PERCEPTRON) Consider the simple network $N = (V, E)$ with $V := \{v_0, v_1, v_2, \ldots, v_n\}$ and $E := \{v_1, v_2, \ldots, v_n\} \times \{v_0\}$. Obviously, we have $\partial := \mathrm{per}(N) = \mathrm{pa}(v_0)$ and $\mathrm{int}(N) = \{v_0\}$. We assume that all neurons are binary: $\Omega_v := \{\pm 1\}$, $v \in V$. The input distribution on $\{\pm 1\}^\partial$ is denoted by p^∂. For each input $\omega = (\omega_1, \ldots, \omega_n)$ (we identify ω_i with ω_{v_i}), the neuron v_0 computes the weighted sum of the input activities according to a vector $\theta = (\theta^1, \ldots, \theta^n) \in \mathbb{R}^n$ of synaptic weights:

$$h_\theta(\omega) \;:=\; \sum_{i=1}^{n} \theta^i\, \omega_i, \qquad \omega \in \{\pm 1\}^\partial.$$

Then it makes a transition to the state $+1$ with probability $f\big(h_\theta(\omega)\big)$, where $f : \mathbb{R} \to]0, 1[$ is the logistic function

$$x \;\;\mapsto\;\; f(x) := \frac{1}{1 + \exp(-x)}.$$

Thus, for $\theta \in \mathbb{R}^n$ we have the transition kernel $k(\cdot \,|\, \cdot, \theta)$ defined by

$$k(+1 \,|\, \omega, \theta) \;=\; \left(1 + \exp\left(-\sum_{i=1}^{n} \theta^i\, \omega_i\right)\right)^{-1}.$$

This model implies the following parametrization

$$\varphi : \;\; \theta \;\mapsto\; p(\omega \,|\, \theta) := p^\partial(\omega_\partial) k(\omega_0 \,|\, \omega_\partial, \theta).$$

After elementary calculations we get the first fundamental form

$$g_{i,j}(\theta) \;=\; \sum_{\omega \in \{\pm 1\}^\partial} p^\partial(\omega)\, \frac{df}{dx}\big(h_\theta(\omega)\big)\, \omega_i\, \omega_j, \qquad 1 \le i, j \le n,$$

and

$$\frac{\partial I^{loc}\big(v_0; \partial\big)}{\partial\, \theta^i}(\theta) \;=$$

$$\sum_{\omega\in\{\pm1\}^\partial} p^\partial(\omega)\,\omega_i\,\frac{df}{dx}\big(h_\theta(\omega)\big)\left(h_\theta(\omega)-\ln\frac{\sum_{\sigma\in\{\pm1\}^\partial}p^\partial(\sigma)f\big(h_\theta(\sigma)\big)}{\sum_{\sigma\in\{\pm1\}^\partial}p^\partial(\sigma)\Big(1-f\big(h_\theta(\sigma)\big)\Big)}\right). \tag{12}$$

Consider the functional that assigns to each random variable $X:\{\pm1\}^\partial\to]0,1[$ the modulated variable

$$\widetilde{X}\;:=\;X(1-X)\left(\ln\frac{X}{1-X}\;-\;\ln\frac{\mathbb{E}_{p^\partial}(X)}{1-\mathbb{E}_{p^\partial}(X)}\right). \tag{13}$$

This modulation is illustrated in the following picture by the shape of the function $x\mapsto x(1-x)\left(\ln\frac{x}{1-x}\;-\;c\right)$:

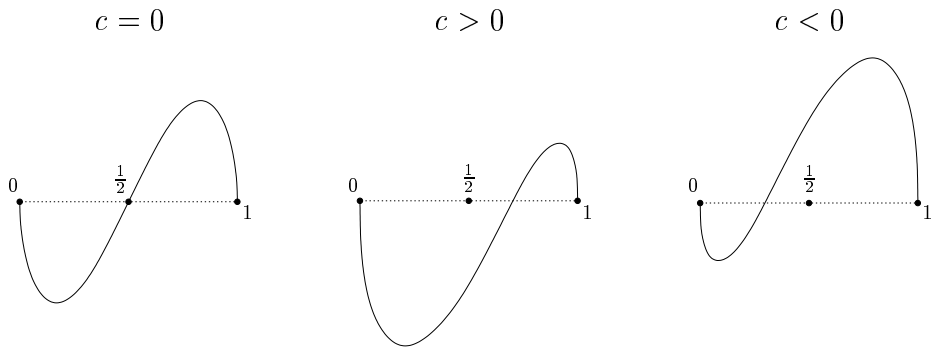

FIG. 4.

Applying the modulation (13) to the variable $X_\theta:=f\circ h_\theta$, (12) simplifies to

$$\frac{\partial I^{loc}\big(v_0;\partial\big)}{\partial\,\theta^i}(\theta)\;=\;\sum_{\omega\in\{\pm1\}^\partial}p^\partial(\omega)\,\omega_i\,\widetilde{X_\theta}(\omega). \tag{14}$$

Thus, one can interprete this part of the gradient to be Hebb-like, where the contribution of the output is not linear but appropriately modulated.

4 Discussion

4.1 More on Linsker's Work

In 1986 Linsker implemented a Hebb-like learning rule in a parametrized feed-forward network which led to the emergence of receptive fields similar to those in the early visual pathway of mammals (Hubel & Wiesel, 1962, 1968). In his description, Linsker considered separately the building blocks of the layered network consisting of two neighbouring layers L and M:

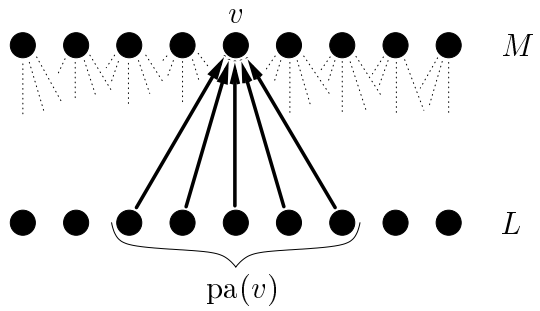

FIG. 5.

Linsker observed that the underlying Hebb-like learning is compatible with the optimization of local information flow (Linsker, 1988, p. 112):

"For more general Hebb-type rules, we found that variance of the output activity was maximized subject to various constraints. This result led us to suggest

that, at least in an intuitive sense, a Hebb rule may act to generate an M cell whose output activity preserves maximum information about the input activities, subject to constraints."

This statement describes in an intuitive sense the optimization of the local interaction of each M-neuron with its parents. According to Theorem 3.8, this optimization can be thought to be driven by the IMI principle. Linsker postulated the *principle of maximum information preservation (Infomax)* which is different from the IMI. According to Linsker's Infomax, the mutual information $I(L; M)$ between the input layer L and the output layer M should be maximized. The Infomax is obtained by induction from the observation that the local information flow is maximized (Linsker, 1988, p. 113):

"The formulation of this principle arose from studying Hebb-type rules and recognizing certain optimization properties to which they lead for single M cells. Once formulated, however, the principle is independent of any particular local algorithm, whether Hebb-related or otherwise, that may be found to implement it."

Linsker's Infomax has been applied to several models, but one important question remains open: Is this principle consistent? The phenomenologically convincing optimization principle that Linsker started with is the *local* one which was implemented by Hebb-like learning rules. Is it possible to recover the local

principle by applying the induced one to the original model? As stated above, the IMI principle has this consistency for the presently considered building block of two layers.

4.2 Some Problems and Comments

(i) Application to general layered networks: In order to apply the IMI optimization to Linsker's model systematically, instead of considering the two-layer building blocks separately, the feed-forward network should be analyzed as a whole. This has only partially been done in the present paper (Proposition 3.6). In particular, the derivation of learning rules in terms of locally adapted stochastic processes that optimize the global stochastic interaction is necessary for establishing the strong local-global connection of interaction in general directed acyclic networks.

(ii) Translation to a dynamical setting: An important property of the IMI principle is its independence from any model specification. In particular, IMI can be applied to arbitrary graphical structures within the setting of random fields (directed, non-directed, or mixed in terms of chain graphs in the sense of Lauritzen, 1996, and Cowell et al, 1999). Nevertheless, in order to develop a dynamical theory of strongly interacting units we considered the temporal aspects of interaction (Ay, 2001b; Ay & Wennekers, 2001). The application of this

dynamical approach to recurrent neural networks is in progress.

(iii) Preference for neural models: Starting with the hypothesis that neural networks are realizations of complex systems in the sense that the elementary units are strongly interacting with each other, it should not only be assumed that *learning* produces high complexity but also that *evolution* generates structural and neurophysiological properties so that the production of high complexity becomes possible. Thus, there should be a preference for intrinsic constraints with sufficient variability, such that learning according to the IMI hypothesis leads to the strong interaction between neurons. On the other hand, too large variability has a negative effect on the generalization ability of learning systems. This leads to the question on the existence of low-dimensional neuro-manifolds that are compatible with the IMI optimization. In connection with the unconstrained optimization, we proved the following theorem (Theorem 3.5 of Ay, 2001a):

Consider n binary neurons, $n \geq 8$. Then there exists an exponential family $\mathcal{E}$ of dimension less than or equal to n^2 such that all distributions with maximal interaction of the neurons can be captured by $\mathcal{E}$.

It is not known how to relate such exponential families to specific neural network models. For instance, one can investigate the manifold of Boltzmann machines with regard to its variability for the IMI optimization.

5 Appendix: Proofs

PROOF OF PROPOSITION 3.3. The lower bound is trivial. We prove the upper bound. Define

$$\mathcal{R} \;:=\; \left\{q \in \bar{\mathcal{P}}(\Omega_V) \;:\; q_\partial = p^\partial,\;\; q_v = p_v \text{ for all } v \in V \setminus \partial\right\} \;\subset\; \mathcal{C}(p^\partial).$$

The set $\mathcal{R}$ can be considered as the intersection of $\bar{\mathcal{P}}(\Omega_V)$ with an affine subspace $\mathcal{V}$ of $\operatorname{aff}\bar{\mathcal{P}}(\Omega_V) \subset \mathbb{R}^{\Omega_V}$ that is given by

$$r \;=\; |\operatorname{supp} p^\partial| - 1 + \sum_{v \in V\setminus\partial} \big(|\Omega_v| - 1\big)$$

linear equations. Of course, $\dim \mathcal{V} \leq \big(|\Omega_V| - 1\big) - r$. Consider the open face of the simplex $\bar{\mathcal{P}}(\Omega_V)$

$$\mathcal{P} \;:=\; \left\{q \in \bar{\mathcal{P}}(\Omega_V) \;:\; \operatorname{supp} q = \operatorname{supp} p\right\}.$$

Then the set

$$\mathcal{S} \;:=\; \mathcal{V} \cap \mathcal{P} \;=\; \mathcal{R} \cap \mathcal{P}$$

is relatively open (i.e., $\mathcal{S}$ is open in $\operatorname{aff}\mathcal{S}$), and $p \in \mathcal{S}$. Furthermore, p locally maximizes the strictly convex restriction $I^V|_{\mathcal{S}}$ of the interaction I^V to $\mathcal{S}$. Thus, p must be an extreme point of $\mathcal{S}$, which is only possible if $\mathcal{S} = \mathcal{V} \cap \mathcal{P} = \{p\}$. This implies

$$\mathcal{V} \;\cap\; \operatorname{aff}\mathcal{P} \;=\; \{p\}. \tag{15}$$

Now we apply the dimension formula

$$\begin{aligned}
|\Omega_V| - 1 &= \dim \operatorname{aff} \bar{\mathcal{P}}(\Omega_V) \\
&\geq \dim \operatorname{aff} \Big(\mathcal{V} \cup \operatorname{aff} \mathcal{P} \Big) \\
&= \dim \mathcal{V} + \big(|\operatorname{supp} p| - 1 \big) - \underbrace{\dim \big(\mathcal{V} \cap \operatorname{aff} \mathcal{P} \big)}_{\overset{(15)}{=} 0} \\
&\geq \big(|\Omega_V| - 1 \big) - r + |\operatorname{supp} p| - 1.
\end{aligned}$$

This gives us the estimate (8). □

Proof of Corollary 3.4. Assume that there is an $\omega \in \operatorname{supp} p^{\partial}$ with

$$|\operatorname{supp} p(\cdot \,|\, X_{\partial} = \omega)| \;>\; 1 + \sum_{v \in V \setminus \partial} \big(|\Omega_v| - 1 \big).$$

Then

$$\begin{aligned}
|\operatorname{supp} p| &= \sum_{\omega \in \operatorname{supp} p^{\partial}} |\operatorname{supp} p(\cdot \,|\, X_{\partial} = \omega)| \\
&> \Big(|\operatorname{supp} p^{\partial}| - 1 \Big) + \left(1 + \sum_{v \in V \setminus \partial} \big(|\Omega_v| - 1 \big) \right).
\end{aligned}$$

This is a contradiction to the estimate (8). □

Proof of Corollary 3.5. This follows directly from Corollary 3.4. □

Proof of Proposition 3.6. We choose a numbering $\{v_1, \ldots, v_m\}$ of $\operatorname{int}(N)$, such that $\operatorname{pa}(v_k) \subset \operatorname{per}(N) \uplus \{v_1, \ldots, v_{k-1}\}$ for all $k = 1, \ldots, m$. This is always possible for directed acyclic graphs. Then, with the chain rule for the entropy

(Cover & Thomas, 1991, p. 21) and the Markov property of p with respect to N (Cowell et al, 1999, p. 74: Theorem 5.14) we get

$$\begin{aligned} H_p(V) &= H_{p^\partial}(\partial) + H_p(v_1 \,|\, \partial) + \cdots + H_p(v_m \,|\, \partial, v_1, \ldots, v_{m-1}) \\ &= H_{p^\partial}(\partial) + H_p\big(v_1 \,|\, \mathrm{pa}(v_1)\big) + \cdots + H_p\big(v_m \,|\, \mathrm{pa}(v_m)\big). \end{aligned}$$

This implies

$$\begin{aligned} I_p^V &= \sum_{v \in V} H_p(v) \;-\; H_p(V) \\ &= \left(\sum_{v \in \partial} H_{p^\partial}(v) - H_{p^\partial}(\partial) \right) \;+\; \sum_{i=1}^{m} \Big(H_p(v_i) - H_p\big(v_i \,|\, \mathrm{pa}(v_i)\big) \Big) \\ &= I_{p^\partial}^{\partial} \;+\; \sum_{v \in \mathrm{int}(N)} I_p\big(v; \mathrm{pa}(v)\big). \end{aligned}$$

□

Proof of Proposition 3.7. We prove this statement by using Proposition 3.6. If p is a local maximizer of $I\big(v; \mathrm{pa}(v)\big)$ for all $v \in \mathrm{int}(N)$ then it also locally maximizes the global interaction I^V in the set of (N, p^∂)-adapted distributions. Thus, we have only to prove the opposite implication. Let $p = p^\partial \otimes \big(\otimes_{w \in \mathrm{int}(N)} K_w\big)$ be a local maximizer of I^V in the set of (N, p^∂)-adapted distributions and assume that there exists a neuron $v \in \mathrm{int}(N)$ such that p is not a local maximizer of $I\big(v; \mathrm{pa}(v)\big)$. Then there exist neighbourhoods $\mathcal{U}_w$ of K_w in $\bar{\mathcal{K}}_w$ with the following properties (i) and (ii):

(i) For all $K'_w \in \mathcal{U}_w$, $w \in \mathrm{int}(N)$, with $p^\partial \otimes \big(\otimes_{w \in \mathrm{int}(N)} K'_w\big) \neq p$ one has

$$I^V_{p^\partial \otimes \left(\otimes_{w \in \mathrm{int}(N)} K'_w\right)} \;\leq\; I_p^V. \tag{16}$$

(ii) There exists a kernel $K_v' \in \mathcal{U}_v$, $K_v' \neq K_v$, such that

$$I_q\big(v; \mathrm{pa}(v)\big) \;>\; I_p\big(v; \mathrm{pa}(v)\big), \tag{17}$$

with

$$q \;:=\; p^{\partial} \otimes \Big(\otimes_{\substack{w \in \mathrm{int}(N) \\ w \neq v}} K_w \;\otimes\; K_v' \Big) .$$

Note that for q and $w \in \mathrm{int}(N)$, $w \neq v$, the following holds:

$$I_q\big(w; \mathrm{pa}(w)\big) \;=\; I_p\big(w; \mathrm{pa}(w)\big). \tag{18}$$

We choose a kernel K_v' that satisfies (17). Applying Proposition 3.6, we get the following contradiction to (16):

$$\begin{aligned} I_q^V \quad &= \quad I_q^{\partial} \;+\; \sum_{\substack{w \in \mathrm{int}(N) \\ w \neq v}} I_q\big(w; \mathrm{pa}(w)\big) \;+\; I_q\big(v; \mathrm{pa}(v)\big) \\ &\overset{(17),(18)}{>} \quad I_p^{\partial} \;+\; \sum_{\substack{w \in \mathrm{int}(N) \\ w \neq v}} I_p\big(w; \mathrm{pa}(w)\big) \;+\; I_p\big(v; \mathrm{pa}(v)\big) \\ &= \quad I_p^V . \end{aligned}$$

The second statement of Proposition 3.7 concerning the "isolated local maximizers" can be proven in the same way. Here, one has to replace in (16) "$\leq$" by "$<$" and in (17) "$>$" by "$\geq$". □

In order to prove Theorem 3.8 we need the following lemma:

LEMMA 5.1. *Let N be a simple network, and let φ be a parametrization of the form (11). Then for each r, the matix $G_r = \left(g_{(r,i)(r,j)}\right)_{i,j}$ only depends on the local parameter vector $\theta^{(r)}$, and the first fundamental form of the Fisher metric is given by*

$$G(\theta) = \begin{pmatrix} G_1(\theta^{(1)}) & & 0 \\ & \ddots & \\ 0 & & G_m(\theta^{(m)}) \end{pmatrix}.$$

PROOF. With the score function

$$\frac{\partial \ln p(\omega \mid \cdot)}{\partial \theta^{(r,i)}}(\theta) \;=\; \frac{\partial \ln k_r(\omega_{v_r} \mid \omega_{\mathrm{pa}(v_r)}, \cdot)}{\partial \theta^{(r,i)}}(\theta^{(r)}),$$

one has

$$\begin{aligned} & g_{(r,i)(s,j)}(\theta) \\ &= \sum_{\omega \in \Omega_V} p(\omega \mid \theta) \, \frac{\partial \ln p(\omega \mid \cdot)}{\partial \theta^{(r,i)}}(\theta) \frac{\partial \ln p(\omega \mid \cdot)}{\partial \theta^{(s,j)}}(\theta) \\ &= \sum_{\omega \in \Omega_V} p(\omega \mid \theta) \, \frac{\partial \ln k_r(\omega_{v_r} \mid \omega_{\mathrm{pa}(v_r)}, \cdot)}{\partial \theta^{(r,i)}}(\theta^{(r)}) \, \frac{\partial \ln k_s(\omega_{v_s} \mid \omega_{\mathrm{pa}(v_s)}, \cdot)}{\partial \theta^{(s,j)}}(\theta^{(s)}). \end{aligned}$$

We consider the case $r \neq s$:

$$
\begin{aligned}
& g_{(r,i)(s,j)}(\theta) \\
&= \sum_{\omega \in \Omega_V} p^{\partial}(\omega_{\partial}) \prod_{\substack{t=1 \\ t \neq r,s}}^{m} k_t(\omega_{v_t} \,|\, \omega_{\mathrm{pa}(v_t)}, \theta^{(t)}) \cdot \ldots \\
&\quad \ldots \left(k_r(\omega_{v_r} \,|\, \omega_{\mathrm{pa}(v_r)}, \theta^{(r)}) \, \frac{\partial \, \ln k_r(\omega_{v_r} \,|\, \omega_{\mathrm{pa}(v_r)}, \cdot)}{\partial \theta^{(r,i)}}(\theta^{(r)}) \right) \cdot \ldots \\
&\quad \ldots \left(k_s(\omega_{v_s} \,|\, \omega_{\mathrm{pa}(v_s)}, \theta^{(s)}) \, \frac{\partial \, \ln k_s(\omega_{v_s} \,|\, \omega_{\mathrm{pa}(v_s)}, \cdot)}{\partial \theta^{(s,j)}}(\theta^{(s)}) \right) \\
&= \sum_{\omega \in \Omega_V} p^{\partial}(\omega_{\partial}) \prod_{\substack{t=1 \\ t \neq r,s}}^{m} k_t(\omega_{v_t} \,|\, \omega_{\mathrm{pa}(v_t)}, \theta^{(t)}) \cdot \ldots \\
&\quad \ldots \frac{\partial \, k_r(\omega_{v_r} \,|\, \omega_{\mathrm{pa}(v_r)}, \cdot)}{\partial \theta^{(r,i)}}(\theta^{(r)}) \, \frac{\partial \, k_s(\omega_{v_s} \,|\, \omega_{\mathrm{pa}(v_s)}, \cdot)}{\partial \theta^{(s,j)}}(\theta^{(s)}) \\
&= \frac{\partial}{\partial \theta^{(r,i)}} \left\{ \frac{\partial}{\partial \theta^{(s,j)}} \left\{ \underbrace{\sum_{\omega \in \Omega_V} p^{\partial}(\omega_{\partial}) \prod_{\substack{t=1 \\ t \neq r,s}}^{m} k_t(\omega_{v_t} \,|\, \omega_{\mathrm{pa}(v_t)}, \theta^{(t)}) \, k_r(\omega_{v_r} \,|\, \omega_{\mathrm{pa}(v_r)}, \cdot) \, k_s(\omega_{v_s} \,|\, \omega_{\mathrm{pa}(v_s)}, \cdot)}_{\equiv 1} \right\}_{\theta^{(s)}} \right\}_{\theta^{(r)}} \\
&= 0.
\end{aligned}
$$

It remains to prove that $g_{(r,i)(r,j)}$ is a function of the local parameter vector $\theta^{(r)}$.

$$
\begin{aligned}
g_{(r,i)(r,j)}(\theta) &= \sum_{\omega \in \Omega_V} p^{\partial}(\omega_{\partial}) \prod_{t=1}^{m} k_t(\omega_{v_t} \,|\, \omega_{\mathrm{pa}(v_t)}, \theta^{(t)}) \cdot \ldots \\
&\qquad \ldots \frac{\partial \, \ln k_r(\omega_{v_r} \,|\, \omega_{\mathrm{pa}(v_r)}, \cdot)}{\partial \theta^{(r,i)}}(\theta^{(r)}) \, \frac{\partial \, \ln k_r(\omega_{v_r} \,|\, \omega_{\mathrm{pa}(v_r)}, \cdot)}{\partial \theta^{(r,j)}}(\theta^{(r)}) \\
&= \sum_{\omega \in \Omega_{\partial}} p^{\partial}(\omega) \sum_{\omega' \in \Omega_{v_r}} k_r(\omega' \,|\, \omega_{\mathrm{pa}(v_r)}, \theta^{(r)}) \cdot \ldots \\
&\qquad \ldots \frac{\partial \, \ln k_r(\omega' \,|\, \omega_{\mathrm{pa}(v_r)}, \cdot)}{\partial \theta^{(r,i)}}(\theta^{(r)}) \, \frac{\partial \, \ln k_r(\omega' \,|\, \omega_{\mathrm{pa}(v_r)}, \cdot)}{\partial \theta^{(r,j)}}(\theta^{(r)}) \\
&= g_{(r,i)(r,j)}(\theta^{(r)}).
\end{aligned}
$$

□

PROOF OF THEOREM 3.8. According to Proposition 3.6 we have

$$\begin{aligned} I^V \circ \varphi &= I^{\partial}_{p^{\partial}} + \sum_{r=1}^{m} I^{loc}\big(v_r; \mathrm{pa}(v_r)\big) \\ &= I^{\partial}_{p^{\partial}} + \sum_{r=1}^{m} \Big(H_{p^{\partial}}\big(\mathrm{pa}(v_r)\big) - H^{loc}\big(\mathrm{pa}(v_r) \,|\, v_r\big) \Big). \end{aligned}$$

Therefore,

$$\frac{\partial (I^V \circ \varphi)}{\partial \theta^{(r,i)}}(\theta) = \frac{\partial\, I^{loc}\big(v_r; \mathrm{pa}(v_r)\big)}{\partial \theta^{(r,i)}}(\theta^{(r)}) = -\frac{\partial\, H^{loc}\big(\mathrm{pa}(v_r) \,|\, v_r\big)}{\partial \theta^{(r,i)}}(\theta^{(r)}).$$

With Lemma 5.1, this completes the proof. □

References

Amari, Shun-ichi (1985). *Differential-Geometric Methods in Statistics*, Lecture Notes in Statistics **28**, Springer-Verlag: Heidelberg.

Amari, Shun-ichi (1998). *Natural gradient works efficiently in learning*, Neural Computation **10** 251-276.

Amari, Shun-ichi (2001). *Information Geometry on Hierarchy of Probability Distributions*, IEEE Trans. IT. **47** 1701-1711.

Amari, Shun-ichi; Nagaoka, Hiroshi (2000). *Methods of Information Geometry*, AMS, Translations of Mathematical Monographs **191**, Oxford University Press.

Attneave, F. (1954). *Informational aspects of visual perception*, Psychological Review **61** 183-193.

Ay, Nihat (2001a). *An Information-Geometric Approach to a Theory of Pragmatic Structuring*, The Annals of Probability (to appear).

Ay, Nihat (2001b). *Information Geometry on Complexity and Stochastic Interaction*, MPI MIS Leipzig, Preprint no. 95, submitted.

Ay, Nihat; Wennekers, Thomas (2001). *Dynamical Properties of Strongly Interacting Markov Chains*, MPI MIS Leipzig, Preprint no. 107, submitted.

Barlow, H. (1989). *Unsupervised learning*, Neural Computation **1** 295-311.

Barlow, H. (2001). *Redundancy reduction revisited*, Network: Comput. Neural Syst. **12** 241-253.

Bell, Anthony J.; Sejnowski, Terrence J. (1995). *An Information-Maximization Approach to Blind Separation and Blind Deconvolution*, Neural Computation **7** 1129-1159.

Cowell, Robert G.; Dawid, A. Philip; Lauritzen, Steffen L.; Spiegelhalter, David J. (1999). *Probabilistic Networks and Expert Systems*, Statistics for Engineering and Information Science, Springer: New York, Berlin, Heidelberg etc.

Chaitin, Gregory J. (1979). *Toward a mathematical definition of "life"*, in R.Levine and M. Tribus, *The Maximum Entropy Formalism*, MIT Press, 477-498.

Chaitin, Gregory J. (2001). *Exploring Randomness*, Discrete Mathematics and Theoretical Computer Science, Springer-Verlag: London, Berlin, Heidelberg.

Cover, Thomas M.; Thomas, Joy A. (1991). *Elements of Information Theory*, Wiley Series in Telecommunications, Wiley-Interscience: New York etc.

Hubel, D.H.; Wiesel, T.N. (1962). *Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex*, Journal of Physiology **160** 106-154.

Hubel, D.H.; Wiesel, T.N. (1968). *Receptive fields and functional architecture of monkey striate cortex*, Journal of Physiology (Lond.), **195** 215-243.

Jost, Jürgen (2000/2001). *Komplexe Systeme*, Lecture given at the University of Leipzig.

Kullback, S.; Leibler, R. A. (1951). *On information and sufficiency*, Annals of Mathematical Statistics **22** 79-86.

Laughlin, S. (1981). *A simple coding procedure enhances a neuron's information capacity*, Z. Naturforsch. **36** 910-912.

Lauritzen, Steffen L. (1996). *Graphical Models*, Oxford Statistical Science Series **17**, Clarendon Press: Oxford.

Linsker, R. (1986). *From Basic Network Principles to Neural Architecture*, Proceedings of the National Academy of Sciences, USA **83** 7508-7512.

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

Linsker, R. (1997). *A local Learning Rule That Enables Information Maximization for Arbitrary Input Distributions*, Neural Computation **9** 1661-1665.

Martignon, L.; Von Hasseln, H.; Grün, S.; Aertsen, A.; Palm, G. (1995). *Detecting higher-order interactions among the spiking events in a group of neurons*, Biological Cybernetics **73** 69-81.

Obradovic, D.; Deco, G. (1998). *Information Maximization and Independent Component Analysis: Is There a Difference?*, Neural Computation **10** 2085-2101.

Rieke, Fred; Warland, David; Ruyter van Steveninck, Rob; Bialek William (1998). *Spikes: Exploring the Neural Code*, Computational Neuroscience, MIT Press: Cambridge, Massachusetts, London, England.

Tononi, Giulio; Sporns, Olaf; Edelmann, Gerald M. (1994). *A measure for brain complexity: Relating functional segregation and integration in the nervous system.* Proc. Natl. Acad. Sci. USA **91** 5033-5037.

Nihat Ay

MPI for Mathematics in the Sciences

Inselstr. 22-26

04103 Leipzig, Germany

E-mail: nay@mis.mpg.de