Temporal Infomax Leads to Almost Deterministic Dynamical Systems

Nihat Ay and Thomas Wennekers[1]

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

Abstract

The well-known Kullback-Leibler divergence of a random field from its factorization quantifies spatial interdependences of the corresponding stochastic elements. We introduce a generalized measure called 'stochastic interaction' that captures also temporal interdependences. Maximization of stochastic interaction in the setting of Markov chains is shown analytically and by simulations to result in an almost deterministic global dynamics, but almost unpredictable single unit activity.

Key words: Markov model; Stochastic interaction; Information maximization;

[1] Tel +49-341-9959558, Fax +49-341-9959555, Email (nay,wenneker)@mis.mpg.de

1 Introduction

Information processing in the brain essentially depends on interrelations between neurons expressed by cooperation and competition in neural assemblies. The generation of spatio-temporal activation patterns, for instance, is at the base of statistical analyses as well as functional theories for neural synchronization and associative propagation of activity [1,6,9,11]. The hypothesis of strong interrelations is contained also in many conceptual approaches to an understanding of first principles for neural cooperation and learning. Information theory provides an appropriate framework for an analysis of such principles [4,7,8,10]. A well-known measure that quantifies relations of interacting units is the so-called *mutual information*: The Kullback-Leibler divergence

$$I(p) \ := \ D(p \,\|\, p_1 \otimes \cdots \otimes p_N) \ = \ \sum_{\nu=1}^{N} H(p_\nu) - H(p) \ , \qquad (1)$$

where $H(\cdot)$ denotes the Shannon entropy and p_ν the ν'th marginal of p, measures the "distance" of p from the set of factorized random fields. It is a natural measure for "spatial" interdependence and a starting point of many approaches to neural complexity. In [2,3] it has been studied from the *information-geometric* point of view, where it is referred to as (*stochastic*) *interaction*. In order to capture also intrinsically temporal aspects of dynamic interaction, I in (1) has been extended by Ay (cf. [4]) to the dynamical setting of Markov chains. In the present paper we consider processes that optimize the temporal version of stochastic interaction. This leads to analytical results concerning the most fundamental feature appearing in strongly interacting stochastic systems, namely, the development of the system dynamics towards determinism. Thereby, a necessary first step is provided for an investigation of learning processes also in more detailed recurrent neural network models.

2 Theoretical Framework and Results

Consider the set $V = \{1, \ldots, N\}$ of binary units with state sets $\Omega_\nu = \{0, 1\}$, $\nu \in V$. For a subsystem $A \subset V$, $\Omega_A := \{0, 1\}^A$ denotes the set of all configurations on A, and $\bar{P}(\Omega_A)$ is the set of probability distributions on Ω_A. Given two subsets A and B, where B is non-empty, $\bar{K}(\Omega_B \,|\, \Omega_A)$ is the set of all Markov transition kernels from Ω_A to Ω_B. If $A = B$, we use the abbreviation $\bar{K}(\Omega_A)$. For a probability distribution $p \in \bar{P}(\Omega_A)$ and a Markov kernel $K \in \bar{K}(\Omega_B \,|\, \Omega_A)$, the *conditional entropy* of (p, K) is defined as

$$H(p, K) \;=\; - \sum_{\omega \in \Omega_A, \omega' \in \Omega_B} p(\omega)\, K(\omega' \,|\, \omega)\, \ln K(\omega' \,|\, \omega)\ . \qquad (2)$$

We set $H(Y \,|\, X) := H(p, K)$ if X and Y are random variables with $\mathrm{Prob}\{X = \omega,\, Y = \omega'\} = p(\omega)\, K(\omega' \,|\, \omega)$ for all $\omega \in \Omega_A$ and $\omega' \in \Omega_B$. In the following we mainly consider the case $A = B = V$ where $H(p, K)$ measures the average uncertainty about the next state of the system given the present state. Assume p is strictly positive. Then H vanishes iff K is deterministic — that is, if the support set, $\operatorname{supp} K(\cdot \,|\, \omega)$, contains only a single possible transition, $|\operatorname{supp} K(\cdot \,|\, \omega)| = 1$, for all ω. H attains its maximal value $N \ln 2$ iff $K(\omega' \,|\, \omega) = 2^{-N}$ for all $\omega, \omega' \in \Omega_V$.

A Markov kernel $K \in \bar{K}(\Omega_V)$ is called *parallel* if there exist kernels $K^{(\nu)} \in \bar{K}(\Omega_\nu \,|\, \Omega_V)$, $\nu \in V$, such that

$$K(\omega' \,|\, \omega) \;=\; \prod_{\nu \in V} K^{(\nu)}({\omega'}_\nu \,|\, \omega), \qquad \text{for all} \quad \omega, \omega' \in \Omega_V. \tag{3}$$

A kernel $K \in \bar{K}(\Omega_V)$ is called *split* if there exist $K^{(\nu)} \in \bar{K}(\Omega_\nu)$, $\nu \in V$, with

$$K(\omega' \,|\, \omega) \;=\; \prod_{\nu \in V} K^{(\nu)}({\omega'}_\nu \,|\, \omega_\nu), \qquad \text{for all} \quad \omega, \omega' \in \Omega_V. \tag{4}$$

The split kernels are a proper subset of the parallel kernels and represent a dynamical version of the factorizable probability distributions. Thus, in analogy to (1) we define the *stochastic interaction* of the units with respect to a distribution $p \in \bar{P}(\Omega_V)$, p strictly positive, and a transition kernel $K \in \bar{K}(\Omega_V)$ as the p-divergence of K from being split: For this purpose, we define the marginal kernels $K_\nu \in K(\Omega_\nu)$, $\nu \in V$, of K by

$$K_\nu(\omega'_\nu \,|\, \omega_\nu) \;:=\; \frac{\sum_{\substack{\sigma, \sigma' \in \Omega_V \\ \sigma_\nu = \omega_\nu,\, \sigma'_\nu = \omega'_\nu}} p(\sigma)\, K(\sigma' \,|\, \sigma)}{\sum_{\substack{\sigma \in \Omega_V \\ \sigma_\nu = \omega_\nu}} p(\sigma)}, \qquad \omega_\nu, \omega'_\nu \in \Omega_\nu. \tag{5}$$

Then, the *stochastic interaction measure* I is defined as

$$I(p, K) \;:=\; \sum_{\nu \in V} H(p_\nu, K_\nu) \;-\; H(p, K)\;, \tag{6}$$

continuously extended to $\bar{P}(\Omega_V) \times \bar{K}(\Omega_V)$ (cf. [4]). Equation (6) generalizes (1) to Markov transitions. $I(p, K)$ is large if the marginal transitions have high entropy, but that of the full transition is low. Supposed the current state $\omega \in \Omega_V$ is known, this corresponds with high predictability of the next global state, but, conversely, not much information is gained from knowledge about the states of single units, ω_ν. The predictability and, hence, degree of determinism of systems that optimize I is further characterized by the following Theorem:

Theorem 1: Consider a probability distribution $p \in \bar{P}(\Omega_V)$ and a transition kernel $K \in \bar{K}(\Omega_V)$. If (p, K) is a local maximizer of the interaction measure I then for all $\omega \in \operatorname{supp} p$ one has $|\operatorname{supp} K(\cdot\,|\,\omega)| \;\leq\; 1 + N$.

Theorem 1 presents a restriction to binary units of a more general Theorem proved in [5]. Note, that the expression $|\operatorname{supp} K(\cdot\,|\,\omega)|$ counts the number of transitions with non-vanishing probability from an arbitrary state $\omega \in$ supp p to its target states. Since there are exponentially many possible target states, namely 2^N, the estimate in Theorem 1, which is only linear in N, provides a strong upper bound on the number of transitions in strongly interacting systems. This has a direct implication for the entropy in such systems:

Corollary 2: In the situation of Theorem 1: $H(p, K) \;\leq\; \ln(1 + N)$.

3 Simulations

Consider a Markov chain $X_n = (X_{\nu,n})_{\nu \in V}$, $n = 0, 1, 2, \ldots$, given by an initial distribution $p \in \bar{P}(\Omega_V)$ and a kernel $K \in \bar{K}(\Omega_V)$. A probability distribution $p \in \bar{P}(\Omega_V)$ is called *stationary* with respect to $K \in \bar{K}(\Omega_V)$ iff $\sum_{\omega \in \Omega_V} p(\omega)\, K(\omega' \,|\, \omega) = p(\omega')$ for all $\omega' \in \Omega_V$. In the present section we consider only Markov chains that are induced by parallel kernels and corresponding stationary distributions. In that case (cf. [4]):

$$I(p, K) \;=\; \sum_{\nu \in V} \Big(H(X_{\nu,n+1} \,|\, X_{\nu,n}) - H(X_{\nu,n+1} \,|\, X_n) \Big). \qquad (7)$$

In (7) the term $H(X_{\nu,n+1} \,|\, X_{\nu,n}) - H(X_{\nu,n+1} \,|\, X_n)$ measures the reduction of uncertainty about unit ν's next state by the additional knowledge of the current states of all other units. In the following we show representative simulations of small complex systems with strong interaction. The simulations implement the usual Markov dynamics on a set of N binary units together with a random search scheme to optimize the stochastic interaction: In every simulation step I is computed for a given parallel Markov kernel and induced stationary probability distribution, and starting from initial random values that kernel is iteratively perturbed such that I increases (cf. [5]).

Figure 1 displays an optimized system for $N = 3$ that is typical also for larger systems. The approached entropies and interaction were $H(p, K) = 0.125$, $H(p_1, K_1) = 0.620$, $H(p_2, K_2) = 0.642$, $H(p_3, K_3) = 0.666$, $I(p, K) =$

1.803. The interaction, I, comes near to its theoretical upper bound of $N \ln 2 \approx 2.07$ because the marginal kernel entropies are all near their maximum of $\ln 2 \approx .69$, and the full kernel entropy is very small (cf. (6)). This indicates that given the current state ω, the next state can be quite reliably predicted, but in contrast, not much information is gained from knowledge about the state of any individual unit alone. The Markov matrix in Fig. 1 (circle area represents transition probability) and the derived state transition graph (node-labels indicate states, ω, edge-labels transition probabilities) exhibit the determinism even more clearly: Most columns in the Markov matrices have only a single entry of probability 1 and, accordingly, most nodes in the transition graphs have just one outgoing edge.

*********** Figure 1 somewhere here ***********

Nonetheless, the transition graph reveals that the dynamics consists of an ergodic component comprising two cycles of deterministic transitions nested by a *branching state* (state 001). Also visible is a *transient state*, (110), which once left is never occuppied again – thus, $p(\omega) = 0$ for transient states. Theorem 1 bounds the maximum number of outgoing transitions to $\leq N{+}1$ for nodes with $p(\omega) > 0$ (the ergodic component). In a large number of simulations with N up to 8, we observed an increasing complexity in the transition graph structure, but seldomly ergodic branching nodes with more than 2 outgoing transitions, and never ergodic nodes with $> N{+}1$ transitions. Transient states, in contrast, can be deterministic as well as project to an arbitrary number of targets up to all possible ones (e.g., 110 in Fig. 1). Branching nodes and nested loops result in activity patterns that switch randomly between repetetive deterministic configuration sequences of various lengths (Fig. 1, lower plot).

*********** Figure 2 somewhere here ***********

Figure 2 shows distributions of $H(p)$ and $H(p, K)$ for $N = 4$ and 7. $H(p, K)$ is always small compared to the theoretical upper bound $N \ln 2$. So, the systems are almost deterministic. The ergodic component as in Fig. 1 consists of nested loops, but the graph structure gets increasingly complex and cycles long with system size (not shown). Also, the size of the transient component increases. Therefore, $H(p)$ falls below its maximal value of $N \ln 2$ for larger N.

Overall, Theorem 1 and the simulation results show that the maximization of stochastic interaction leads to Markov chains with highly deterministic global dynamics, but randomness in single unit activity. Future work will address stochastic interaction in more realistic neural network models.

References

[1] M. Abeles, Corticonics: Neural circuits of the cerebral cortex (Cambridge University Press, Cambridge, 1991).

[2] S.-I. Amari, Information Geometry on Hierarchy of Probability Distributions, IEEE Transactions on Information Theory 47 (2001) 1701–1711.

[3] N. Ay, An Information-Geometric Approach to a Theory of Pragmatic Structuring, The Annals of Probability 30 (2002) 416–436.

[4] N. Ay, Information Geometry on Complexity and Stochastic Interaction, Submitted (2002).

[5] N. Ay and T. Wennekers, Dynamics of strongly interacting Markov chains, Submitted (2002).

[6] R. Eckhorn, Neural mechanisms of scene segmentation: Recordings from the visual cortex suggest basic circuits for linking field models, IEEE Transactions on Neural Networks 10 (1999) 464–479.

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

[8] F. Rieke, D. Warland, R. Ruyter van Steveninck and W. Bialek, Spikes: Exploring the Neural Code (MIT Press, Cambridge, 1998).

[9] W. Singer and C.M. Gray, Visual feature integration and the temporal correlation hypotheses, Annual Review of Neuroscience 18 (1995) 555–586.

[10] G. Tononi, O. Sporns and G.M. Edelman, A measure for brain complexity: Relating functional segregation and integration in the nervous system, Proceedings of the National Academy of Sciences USA 91 (1994) 5033–5037.

[11] T. Wennekers and G. Palm, Cell Assemblies, Associative Memory and Temporal Structure in Brain Signals, in: R. Miller, ed., Time and the Brain (Harwood Academic Publishers, 2000) 251–273.

Nihat Ay studied Mathematics at the Ruhr-University Bochum and obtained a Ph.D. in Mathematics from the University of Leipzig in 2001. He currently works at the Max-Planck-Institute for Mathematics in the Sciences in Leipzig on information geometry and its applications in complex adaptive systems.

Thomas Wennekers studied physics at the University of Düsseldorf and computer science at the University of Ulm, where he received a Ph.D. in 1998. He does research at the Max-Planck-Institute for Mathematics in the Sciences in Leipzig in the fields of computational neuroscience and brain theory.

Figure Legends

Fig. 1: Markov matrix (upper left), state transition graph (right), and sample activity (bottom) for a system of $N = 3$ strongly interacting units.

Fig. 2: Left: $H(p)$ and $H(p, K)$ for a series of simulations with $N = 4$ and $N = 7$. A theoretical upper bound for I and $H(p)$ is $N \ln 2 \approx 2.77$ and 4.85, respectively.

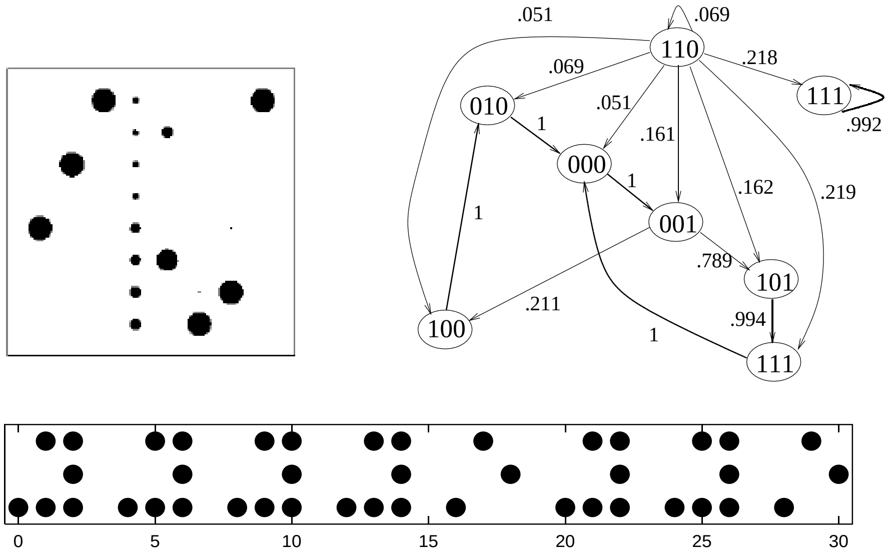

Fig. 1. Ay and Wennekers

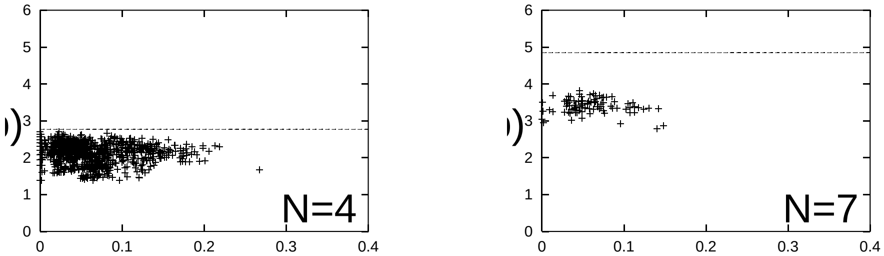

Fig. 2. Ay and Wennekers