

Finite state automata resulting from temporal information maximization

Thomas Wennekers^{1,2,3} and Nihat Ay^{3,4,5}

¹Centre for Theoretical and Computational Neuroscience
University of Plymouth, Plymouth PL4 8AA, United Kingdom
Phone: +44(0)1752-23-3593, Fax: +44(0)1752-23-3349
Email: Thomas.Wennekers@plymouth.ac.uk

² Institute for Neuroinformatics
Ruhruniversity Bochum, 44780 Bochum, Germany

³ Max Planck Institute for Mathematics in the Sciences
Inselstrasse 22–26, 04103 Leipzig, Germany

⁴ Intitute for Mathematics
Friedrich Alexander University Erlangen-Nuremberg
Bismarckstrasse 1 1/2, 91054 Erlangen, Germany
Email: ay@mi.uni-erlangen.de

⁵ Santa Fe Institute
1399 Hyde Park Road, Santa Fe, NM 87501, USA

Abstract

We extend Linker’s Infomax principle for feedforward neural networks to a measure for stochastic interdependence that captures spatial *and* temporal signal properties in recurrent systems. This measure — “stochastic interaction” — quantifies the Kullback-Leibler divergence of a Markov chain from a product of split chains for the single unit processes. For unconstrained Markov chains, the maximization of stochastic interaction also called “Temporal Infomax”, has been previously shown to result in almost deterministic dynamics. The present work considers Temporal Infomax on constrained Markov chains, where part of the units are clamped to prescribed stochastic processes providing external input to the system. Surprisingly, Temporal Infomax in that case leads to finite state automata, either completely deterministic or at most weakly non-deterministic. Transitions between internal states of these systems are almost perfectly predictable given the complete current state and the input, but the activity of each single unit alone is virtually random. The results are demonstrated by means of computer simulations and confirmed analytically. We furthermore relate them to experimental data concerning the correlation dynamics and functional connectivities observed in multiple electrode recordings.

1 Introduction

A fundamental question in computational neuroscience asks for the nature of codes employed by cortical neurons (Dayan & Abbott, 2001; Abbott & Sejnowski, 1999). Experiments suggest a considerable interaction of neurons already on the level of spikes, for instance, expressed by spatio-temporal correlations in multiple unit recordings (Abeles et al., 1993a; Eckhorn, 1990; Rieke et al, 1998; Singer & Gray, 1995). Such correlations have been a matter of intensive theoretical and conceptual research, cf., e.g., Abeles (1990); Abeles et al. (1993b); Aertsen (1993); Gerstein et al. (1989); Palm & Aertsen (1986); Wennekers et al. (2003).

A well-known measure that quantifies spatial relations of interacting units is the *multi-information* (see Studený M. & Vejnarova J. (1998)) shared among the units, which is called *mutual information* in the case of two units (see Cover & Thomas (1991)). Multi-information can be expressed in terms of the Kullback-Leibler divergence as

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

In (1), $H(\cdot)$ denotes the usual Shannon entropy and p_ν the ν ’th marginal of p . $I(p)$ measures the “distance” of p from the factorized distribution $p_1 \otimes \cdots \otimes p_N$. It is a natural measure for “spatial” interdependence of N stochastic units and a starting point of many approaches to neural coding and complexity, e.g., Martignon et al. (1995, 2000); Nakahara & Amari (2002); Rieke et al (1998); Sporns et al. (2000); Tononi et al. (1994); Ay (2002); Wennekers & Ay (2003).

Linsker (1986a,b,c), for instance, considered layered feedforward neural systems which maximize the mutual information between (stationary) input and output probability distributions. These works revealed surprising relations between information maximization

and the spatial structure of receptive fields in the visual system. Recent experiments in that direction suggest that individual neurons may even adapt dynamically to maximize their information transfer with respect to a given stimulus ensemble (Fairhall et al., 2001).

As a fundamental concept regarding neural coding the Infomax principle is still being discussed and developed further towards confined models for the development and functional significance of receptive fields, see e.g., Abbott & Sejnowski (1999); Barlow (2001); Bell & Sejnowski (1995); Li & Atick (1994); Penev & Atick (1996). In addition, theoretical work revealed a close relation between Information Maximization on one hand and principal or independent component analysis on the other. This puts Linsker's Infomax in row with projection techniques for high-dimensional data analysis (Cutler & Breiman, 1994; Hertz et al., 1991; Bell & Sejnowski, 1995; Lee et al., 2000). For these reasons Linsker's Infomax principle can be seen an important guiding principle in computational neuroscience and neuroinformatics.

Neural systems, however, are in general recurrently connected and non-stationary, both features not reflected by most classical information-based approaches to neural complexity. It is therefore of interest to study information maximization also in more general settings. In order to capture intrinsically temporal aspects of dynamic interactions in recurrent networks, the measure I in (1) has been extended by Ay (2001) to the dynamical setting of Markov processes, where it is referred to as (*stochastic*) *interaction*. In a previous paper (Ay & Wennekers, 2003) we have shown that the optimization of stochastic interaction in Markov chains leads to globally almost deterministic dynamical systems, where nonetheless every single unit generates virtually random activity as characterized by a high entropy. That work neglected external input into the systems under study. The present paper therefore investigates the more interesting case of Markov chains, where a part of the system is clamped to prescribed stochastic processes, but only the internal dynamics is optimized towards large stochastic interaction. Surprisingly, Markov processes that optimize stochastic interaction under this input constraint turn out to be finite state automata, where the internal dynamics is driven by the external input through complex, almost deterministic global state sequences, but again, single unit activity is virtually random.

To demonstrate and explain these phenomena the paper is organized as follows: The next section introduces the basic formalism of constrained Markov chains and generalizes Shannon-entropy and mutual information to the spatio-temporal dynamics of Markov chains. Section 3 presents detailed simulations of small example systems with numerically maximized stochastic interaction under the constraint that part of the systems follows prescribed stochastic processes. Section 4 afterwards explains the basic features of strongly interacting systems observed in the simulations analytically. In especially, upper bounds for the amount of order and entropy in certain optimized systems are derived mathematically. Proofs of the stated theorems are sketched in the appendix. The paper closes with a discussion, which relates our results to experiments concerning spatio-temporal correlations in biological neural ensembles.

2 Temporal Infomax on Constrained Markov Chains

Consider a set $V = \{1, \dots, N\}$ 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 restricted to A , and $\bar{\mathcal{P}}(\Omega_A)$ is the set of probability distributions on Ω_A . Given two subsets A and B , where B is non-empty, $\bar{\mathcal{K}}(\Omega_B | \Omega_A)$ is the set of all Markov kernels from Ω_A to Ω_B . In the case $A = B$ we use the abbreviation $\bar{\mathcal{K}}(\Omega_A) = \bar{\mathcal{K}}(\Omega_A | \Omega_A)$.

For a probability distribution $p \in \bar{\mathcal{P}}(\Omega_A)$ and a Markov kernel $K \in \bar{\mathcal{K}}(\Omega_B | \Omega_A)$ we define a *Markov transition* as the pair (p, K) and the *conditional entropy* of (p, K) as

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

$H(p, K)$ defined this way is a natural extension of the Shannon-entropy to Markov transitions, because $-\ln K(\omega' | \omega)$ in (2) is the information content of an individual state transition supposed ω is known and $K(\omega' | \omega)p(\omega)$ is the probability for that transition. Thus, Equation (2) measures the average information generated by the Markov transition (p, K) just as the Shannon entropy measures the average information contained in a stationary probability distribution p : $H(p) = -\sum_{\omega} p(\omega) \ln p(\omega)$.

Note that a Markov transition is not the same as a Markov chain. It describes just a single transformation step from a probability distribution over a set of states Ω_A to a probability distribution over a possibly different set Ω_B . However, if $\Omega_A = \Omega_B$, and p is a stationary probability distribution of the kernel K then a Markov transition induces a Markov chain in a natural way.

The conditional entropy has been used by Ay (2001) to generalize (1) from stationary probability distributions to stochastic processes. This extension enables the definition of a divergence or distance between two Markov transitions and in especially, of a measure for the divergence of a given transition from its corresponding product of marginal transitions for the individual units. For that purpose, we define the marginal kernels $K_\nu \in \mathcal{K}(\Omega_\nu)$, $\nu \in V$, of a kernel $K \in \mathcal{K}(\Omega_V)$ by

$$K_\nu(\omega'_\nu | \omega_\nu) := \frac{\sum_{\substack{\sigma, \sigma' \in \Omega_V \\ \sigma_\nu = \omega_\nu, \sigma'_\nu = \omega'_\nu}} p(\sigma) K(\sigma' | \sigma)}{\sum_{\substack{\sigma \in \Omega_V \\ \sigma_\nu = \omega_\nu}} p(\sigma)}, \quad \omega_\nu, \omega'_\nu \in \Omega_\nu. \quad (3)$$

Equation (3) projects the full kernel $K(\sigma' | \sigma)$ defined on the whole state space to a kernel $K_\nu(\omega'_\nu | \omega_\nu)$ for only unit ν . Clearly, in (3) the expression $p(\sigma) K(\sigma' | \sigma)$ is the probability that the system is in state σ and transits to σ' . Thus, summing over all states σ, σ' with unit ν clamped to ω_ν and ω'_ν respectively, gives the total probability for transitions where unit ν is in state ω_ν before and in state ω'_ν after the transition, irrespective of the rest of the system. The normalization by $p_\nu(\omega_\nu) := \sum_{\substack{\sigma \in \Omega_V \\ \sigma_\nu = \omega_\nu}} p(\sigma)$ in (3) ensures that K_ν is a proper Markov kernel, i.e., $\sum_{\omega'_\nu \in \Omega_\nu} K_\nu(\omega'_\nu | \omega_\nu) = 1$ for all $\omega_\nu \in \Omega_\nu$. In fact, p_ν is the marginal probability distribution for unit ν . Further, the pairs (p_ν, K_ν) , $\nu = 1, \dots, N$ are the marginal Markov transitions of the transition (p, K) . Note that the marginal transition kernels are defined in (3) only for kernels $K \in \mathcal{K}(\Omega_V)$, and not for more general kernels in $\mathcal{K}(\Omega_B | \Omega_A)$. The reason is that we will mainly consider recurrent systems, where

the states before and after the transition refer to the same set of cells V . The definition of the conditional entropy in (2), on the other hand, can also be applied to systems where these sets are different, for instance, to feedforward models where they are disjoint.

The *stochastic interaction measure* of K with respect to p is defined as

$$I(p, K) := \sum_{\nu \in V} H(p_\nu, K_\nu) - H(p, K), \quad (4)$$

with values in the range $[0, \sum_{\nu \in V} \ln |\Omega_\nu|]$, because the minimum of $H(p, K)$ is zero for deterministic systems and the maximum entropy of a single unit with $|\Omega_\nu|$ states is $\ln |\Omega_\nu|$. Evidently, for N binary units the maximal interaction is $N \ln 2$ (or N “bits” if we had chosen the dual instead of the natural logarithm in the definition of the conditional entropy). Comparison with (1) shows that (4) has the form of a Kullback-Leibler divergence and generalizes the usual mutual information to Markov transitions. It measures the divergence of (p, K) from the product of its marginal transitions, thereby indicating how much (p, K) deviates from a product of independent single unit transitions, or, in other words, how strong the units in (p, K) “interact” stochastically. In especially, observe that $I(p, K)$ is particularly large if the marginal transitions have high entropy, but that of the full transition is low. Then, supposed the current state $\omega \in \Omega_V$ is known, the next global state is predictable with high confidence, but, conversely, not much information is gained from knowledge about single units, ω_ν . We call such systems “strongly interacting” and study some of their properties in the sequel.

For that purpose we consider Markov chains $X_n = (X_{\nu, n})_{\nu \in V}$, $n = 0, 1, 2, \dots$, given by an initial distribution $p_0 \in \bar{\mathcal{P}}(\Omega_V)$ and a kernel $K \in \bar{\mathcal{K}}(\Omega_V)$. We further restrict attention basically to *parallel* Markov chains. A Markov kernel $K \in \bar{\mathcal{K}}(\Omega_V)$ is called *parallel* if there exist kernels $K^{(\nu)} \in \bar{\mathcal{K}}(\Omega_\nu | \Omega_V)$, $\nu \in V$, such that

$$K(\omega' | \omega) = \prod_{\nu \in V} K^{(\nu)}(\omega'_\nu | \omega), \quad \text{for all } \omega, \omega' \in \Omega_V. \quad (5)$$

Given the current global state ω each parallel kernel $K^{(\nu)}$ in (5) determines the next state of only a single unit ν independent of transitions in other units. Therefore the kernels can be termed “local” and the global transition is of product form similar as for independent stationary probability distributions. In contrast, general kernels represent mappings between arbitrary global states. A source state ω can specifically target on arbitrary subsets of other global states ω' . The state transition of a certain unit then depends on the simultaneous transitions of other units, i.e., on “non-local” information. As a consequence, parallel Markov chains are a more natural assumption in neural modeling than general Markov chains, because the activity of a neuron is determined only by its own input and internal dynamics, not by the simultaneous activity of other cells. Interaction takes place only through recurrently distributed activity. Therefore, each unit ν in (5) is defined by an individual kernel $K^{(\nu)}$.

We previously considered parallel Markov chains where $I(p, K)$ was numerically maximized under no further constraint regarding K (Ay & Wennekers, 2003). We call this approach *Temporal Infomax*. The optimized, strongly interacting chains were shown to be globally almost deterministic, but the firing of individual units was largely random

and unpredictable. Strongly interacting Markov chains turned out to be representable in state space Ω_V by complex but systematically structured transition graphs consisting of a fraction of transient trajectories as well as sets of attracting nested loops which correspond with almost deterministic repetitive firing patterns of various lengths (cf. Fig. 1). Because these studies did not consider external input into the system, the observed globally deterministic/locally random activity patterns can be envisaged as intrinsic modes of activity in strongly interacting stochastic systems.

The present work, in contrast, considers Markov chains which maximize $I(p, K)$ under the additional constraint that the kernels of a subset $\partial \subset V$ of units are fixed during the optimization process. We call the set of units ∂ the *periphery* of the system and the set $V \setminus \partial$ the *interior* or *internal units*. The peripheral units serve as input units for the rest of the network, but are assumed to be independent of the interior. The latter assumption can be relaxed, i.e., in principle we could also let the peripheral units depend in some way on the activity in the interior, but we do not consider this case in the present work.

Observe further, that the processes prescribed on the periphery are supposed to represent the activity evoked on the sensory surface of the brain by some real world process. Those may comprise spatio-temporal regularities and correlations of quite arbitrary nature. In especially, any stationary probability distribution should be possible for units clamped on the periphery. These reflect a stationary stimulus ensemble driving the peripheral units as in Linsker's original setting. In the setting of Markov chains a sequence of random samples from a stationary distribution corresponds to a memoryless Markov chain, where $K(\omega | \omega') \equiv K(\omega)$, that is, the transition kernel does not depend on the second argument, which represents the one-step memory of the chain. In that case X_{t+1} is (temporally) independent of X_t for all t , but all X_t are drawn from the same probability distribution $p(\omega) = K(\omega)$. If we now could factorize K into a parallel kernel, see (5), then apparently the peripheral units were all (spatially) independent, excluding any kind of correlations. This is clearly a too restrictive model for external stimuli. Therefore, in the sequel we only restrict the internal units to parallel chains but use general Markov chains for the peripheral units. Formally, we write $\omega = (z, a)$ for $\omega \in \Omega_V$, $z \in \Omega_{V \setminus \partial}$, $a \in \Omega_\partial$. If $K^{(\nu)}(\omega'_\nu | \omega) = K^{(\nu)}(\omega'_\nu | z, a)$ denotes the parallel kernels of the internal units and $K^\partial(a' | a)$ the general kernel on the periphery, then for all $\omega, \omega' \in \Omega_V$ the global state transition probabilities are given by

$$K(\omega' | \omega) = K(z', a' | z, a) = K'(z' | z, a) K^\partial(a' | a) \quad (6)$$

$$= \left[\prod_{\nu \in V \setminus \partial} K^{(\nu)}(\omega'_\nu | z, a) \right] K^\partial(a' | a) . \quad (7)$$

For later use, we abbreviated the Markov kernel in square brackets in (7) as $K'(z' | z, a)$ in (6). From (7) it again becomes obvious that for fixed (z, a) the ω'_ν , $\nu \in V \setminus \partial$, and $a' \in \Omega_\partial$ are mutually independent since (7) is of product form.

3 Examples for strongly interacting systems

This section presents some exemplary simulations of strongly interacting Markov chains with various peripheries. These simulations numerically optimize the stochastic interaction measure $I(p, K)$ in (4) for kernels of the form (7).

3.1 Simulation procedures

The simulations displayed in the following sections implement the usual Markov dynamics on a set of N binary units to generate sample trajectories. In addition, a random search scheme is implemented to optimize the stochastic interaction of the Markov chains. This optimization process is completely independent of the trajectory generation, since $I(p, K)$ can be computed from any given kernel K and probability distribution p alone. The sample trajectories displayed in subsequent figures, thus, serve only for visualization of the network dynamics.

Details of the optimization are described in Ay & Wennekers (2003). In brief, in every time-step during optimization, the interaction measure, $I(p, K)$, is computed with respect to an induced stationary probability distribution p of a stored Markov kernel K . A stationary distribution p is determined by solving the equation $Kp = p$ in every step. Usually we start from ergodic Markov chains, where p is unique. If during the optimization a Markov chain becomes non-ergodic we select an arbitrary one of the solutions of $Kp = p$. Starting from initial random values the kernel is then iteratively perturbed such that I increases. In contrast to Ay & Wennekers (2003), however, the optimization here is not unconstrained, but the stored Markov kernels have the form (7), where only the values $K^{(\nu)}(\omega'_\nu | z, a)$, $\nu \in V \setminus \partial$, $\omega'_\nu \in \Omega_\nu = \{0, 1\}$, $z \in \Omega_{V \setminus \partial}$, $a \in \Omega_\partial$ are perturbed during optimization. The peripheral kernels $K^\partial(a' | a)$ are chosen to be independent of the internal units and kept fix during optimization. We consider different choices for the peripheral kernels below, e.g., kernels with all entries equal, randomly initialized kernels, or special deterministic kernels. These kernels will be defined were used.

After convergence, state transition graphs are constructed from the resulting full Markov kernels and plotted using the public domain software *dot*¹. Simulation programs were implemented using the simulation environment **Felix** written by one of the authors (T.W.) and run on various UNIX/Linux platforms. Because the optimization of I is algorithmically complex, simulations are restricted to small N .

3.2 Unconstrained optimization

Figure 1 displays an example of a system with $N = 3$ units optimized with *no* units clamped, i.e., an empty periphery ∂ . Such systems have been studied in detail in Ay & Wennekers (2003). Since important for the subsequent discussion of constrained optimized Markov chains and in order to introduce a couple of basic definitions and phrases, we briefly summarize the main properties of unconstrained chains: Figure 1A shows the

¹The program *dot* is part of the Graph Drawing Package *graphviz* from AT&T and Lucent Bell Labs available under <http://www.research.att.com/sw/tools/graphviz>.

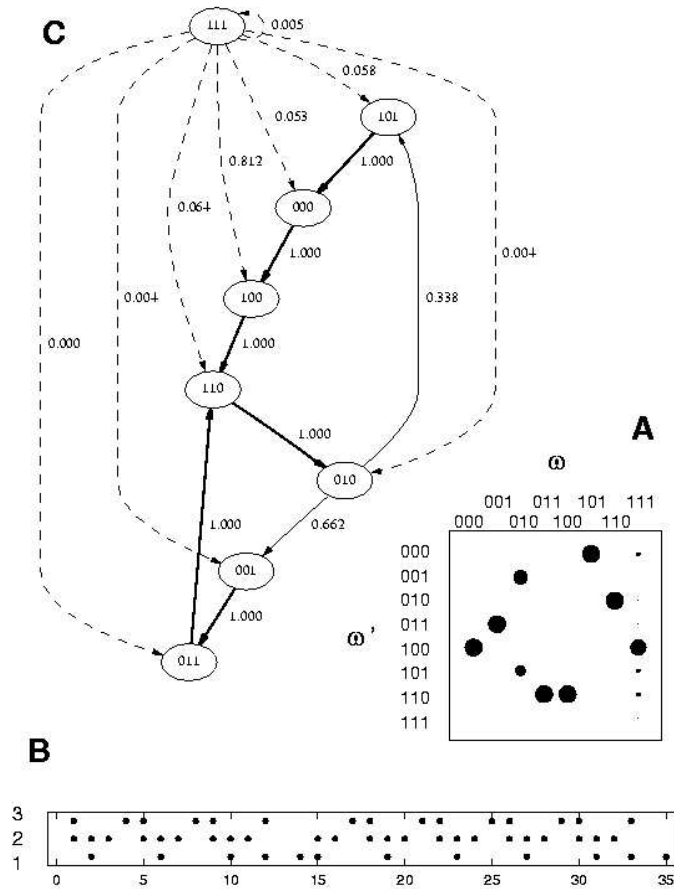


Figure 1: An example for unconstrained optimization using $N = 3$ units. A shows the optimized Markov matrix, where dot-size indicates transition probability. B displays a sample trajectory as a raster plot over time; dots in B correspond with an output of 1. C is the state transition graph representing the matrix in A. Node labels denote states, ω , and edge labels transition probabilities. Observe the almost deterministic asymptotic state transitions (bold edges). State 111 is ‘transient’ and 010 a ‘branching state’.

Markov matrix $K(\omega'|\omega)$ of an optimized system as a dot-display, where dot-size indicates the transition probability. Most columns of the matrix reveal only a single possible transition. This indicates that the respective transitions are deterministic, that is, occur with probability 1 if the source state ω is given. However, there are two exceptions, states $\omega = 111$ and 010. State 111 actually is a *transient* state: It has outgoing transitions to some (here, all) other states, but none of the other states projects back to it. Therefore, once left, state 111 is never occupied again, such that the stationary probability for observing that state in sample trajectories is zero. On the other hand, state 010 is what we call a *branching state*. As Fig. 1C shows, which redisplay the matrix in A as a state transition graph, state 010 is part of two nested loops of states with deterministic transitions between nodes (deterministic transitions are plotted as bold edges in the figure). Only state 010 has two outgoing – and therefore non-deterministic – transitions. In fact, the nested loop structure represents the attractor of the dynamics reached asymptotically

after transients have died out. Activity flows deterministically along consecutive states of the individual loops, but at state 010 it can switch randomly between two possible targets leading back to one or the other deterministic sequence of states. Accordingly, sample trajectories of the dynamics are characterized by randomly interleaved sequences of repetitive deterministic firing patterns as shown in Fig. 1B. The nested loop attractor represents the *ergodic component* of the dynamics; states in the ergodic component have a positive stationary probability $p(\omega)$.

The example in Fig. 1 is also characteristic for larger systems. In contrast to arbitrary Markov chains, where transitions from all to all states are generically possible, the dynamics of strongly interacting chains is confined to a core of nested deterministic subsequences of states linked by branching nodes and augmented by a set of transient states. As we have proved formally in Ay & Wennekers (2003) the number of outgoing edges in branching states is strictly bounded by $N + 1$ as compared to a total of 2^N possible transitions. Simulations reveal that the number is usually even smaller than this theoretical upper bound. Consequently, strongly interacting Markov chains are almost deterministic.

3.3 Example with clamped periphery

Figure 2 shows an example system comprising $N = 4$ units, where two of the units have been clamped to a Markov chain with equal transition probabilities between peripheral states. Figure 2A displays the respective peripheral kernel K^∂ and Fig. 2B the full Markov matrix $K(\omega' | \omega) = K(z', a' | z, a)$. Here, as well as throughout the paper, we assume that units are counted from left to right in binary representations of states z , a , or $\omega = (z, a)$; we do not print the state representations in plots of Markov kernels in later figures.

Now, what is the impact of the clamped periphery on the optimized Markov chain? Clearly, the most prominent difference between the kernels in Fig. 2B and Fig. 1A is that the columns in Fig. 2B do not reveal just one, but four entries (with probabilities summing up to 1, since K is a Markov kernel). These entries, however, are grouped into blocks, as indicated in the figure, and all transitions for one source state (z, a) target in exactly one of the blocks. Furthermore, a closer inspection shows that the blocks are uniquely characterized by the internal states, z , z' , whereas the peripheral states a , a' only indicate the precise location inside each block. Thus, given a source state z and a peripheral state a , the next *internal* target state z' is uniquely defined. In other words, the internal state transition kernel $K'(z' | z, a)$ as defined in (6) is deterministic. The next peripheral state, of course, is random, and indeed independent of the internal state, because by assumption it is completely governed by $K^\partial(a' | a)$ and the current peripheral state a . Thus, starting from some internal initial state, the peripheral Markov dynamics - viewed as input - drives the internal subsystem through deterministic state sequences. Nonetheless, because the dynamics on the periphery is random, sample trajectories do not reveal much determinism at a rough look, see Fig. 2C. However, note that at any time in Fig. 2C given the current state (z, a) , the next internal state z' is perfectly predictable supposed K is known.

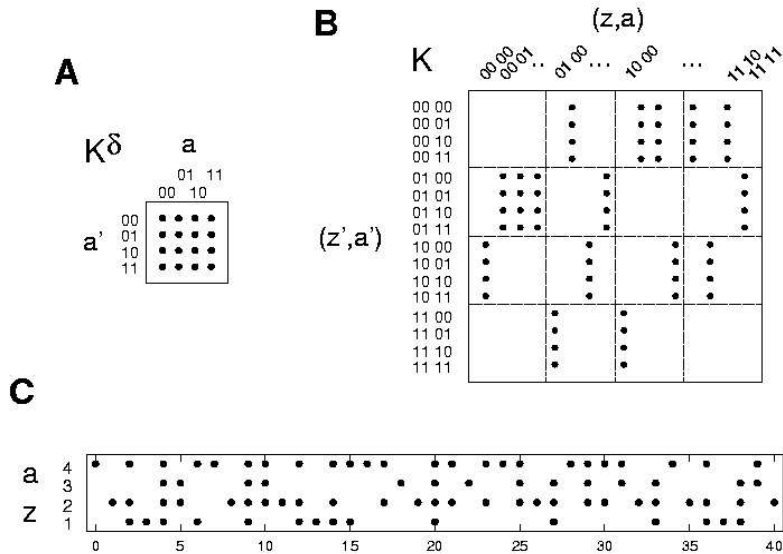


Figure 2: A strongly interacting Markov chain with $N = 4$ units, $|\partial| = 2$ of which clamped to a peripheral chain with equal transition probabilities (0.25) between peripheral states a, a' . Interaction of the system is $I(p, K) = 1.35773 \approx 2 \ln 2$. A) the peripheral kernel $K^\delta(a'|a)$; B) the full Markov kernel $K(\omega'|\omega) = K(z',a'|z,a)$; C) a sample trajectory.

3.4 Strongly interacting Markov chains as automata

Computer science (Hopcroft & Ullman, 1979) defines deterministic finite state automata (DFAs) as a quintuple $M = (Z, \Sigma, \delta, z_0, E)$, where $Z = \{z_1, \dots, z_n\}$ is a finite *set of states* and $\Sigma = \{a_1, \dots, a_m\}$, a finite *alphabet*. The designated state $z_0 \in Z$ is called the *initial state* and $E \subseteq Z$ the set of *final or accepting states*. Operation of the automaton is defined by the *transition table* $\delta : Z \times \Sigma \rightarrow Z$ which maps every pair $(z, a) \in Z \times \Sigma$ to exactly one successor state.

A *word* (over Σ) is any finite sequence (or string) consisting of symbols in Σ . A word x is said to be accepted by a DFA M , iff reading the word symbol by symbol starting from the initial state, application of the respective transition rules leads to a final state in E when the word is read completely. It is known that the set of all words, $L(M)$, that a DFA M accepts is a regular language.

Now, observe that as a finite state automaton, the strongly interacting Markov chain in Fig. 2 provides a total mapping from $\Omega_{V \setminus \partial} \times \Omega_\partial$ to $\Omega_{V \setminus \partial}$. We may therefore identify the internal state space $\Omega_{V \setminus \partial}$ with the state set Z of a DFA, and the peripheral state space Ω_∂ with a set of symbols Σ . The Markov kernel $K'(z'|z, a)$ then corresponds to the transition table of that DFA, and – as all finite state automata – the Markov chain can be represented by a labeled state transition graph as in Fig. 3.

Clearly, for a complete correspondence, we would also have to designate an initial state, z_0 , and accepting states, E . However, z_0 and E merely specify, how an FSA actually decides whether it accepts or rejects a certain input string. We could add equivalent constructs also in our Markov models, e.g., by arbitrarily selecting initial and accepting states and presenting segmented input (finite words), cf. e.g. Wennekers (1998). But

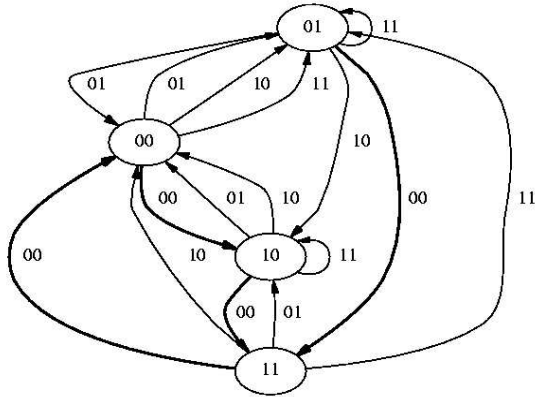


Figure 3: Deterministic finite state automaton corresponding with Fig. 2. Nodes are labeled by internal states $z \in \Omega_{V \setminus \partial}$ and edges by peripheral states $a \in \Omega_{\partial}$. Given the current internal state z and peripheral state a , the automaton predicts the next internal, but not the next peripheral state. Bold edges are for input 00 and can be envisaged as the intrinsic dynamics of the system in the absence of input, cf. section 3.8.

Figure 4: Markov kernels for two strongly interacting Markov chains with $N = 4$, $|\partial| = 2$, and random peripheral kernels K^{∂} . Both matrices reveal non-deterministic transitions (arrows), where a given global state projects to more than one internal target state (i.e. block).

these issues are of secondary importance for the present paper. The main point is that the maximization of stochastic interaction in constrained Markov chains in fact leads to systems characterized by deterministic, input-driven internal state transitions. The kernel K' can then be interpreted as the transition table of a finite state automaton.

3.5 Weakly non-deterministic Markov Chains

The previous section somewhat simplified things: Although most columns in the unconstrained Markov chain in Fig. 1 are deterministic, the example also reveals non-deterministic transitions: transient and branching states have several possible target states. Is there a corresponding phenomenon for constrained Markov chains? The answer is yes and Fig. 4 displays two examples, both for $N = 4$, $|\partial| = 2$, and periphery clamped to randomly initialized kernels (all entries in K^{∂} were set to random values equally distributed on $[0, 1]$; afterwards columns were normalized to sum to 1. Random peripheral kernels are not a necessary condition for non-deterministic transitions. We also observed them with other types or peripheries.)

The left Markov matrix in Fig. 4 reveals just one non-deterministic column for source state $(z, a) = 0010$, the right three of them, all outgoing from internal state $z = 01$. In fact, this state is a transient state with respect to the internal dynamics, because none of the other states projects back to it. Therefore, its stationary occupation probability equals zero. In this case the outgoing transitions of $z = 01$ do not contribute to $H(p, K)$

nor to $I(p, K)$ since the vanishing probabilities $p(01, a) = 0$, $a \in \Omega_\partial$, cancel the respective terms from the sums in (2) and (3). As a consequence, $I(p, K)$ is neutral with respect to changes in probabilities of outgoing edges of $z = 01$. This explains the occurrence of non-deterministic state transitions in Fig. 4B; they correspond with transient states in unconstrained optimized Markov chains. In Fig. 4A the non-deterministic transitions are non-trivial; they relate to branching states.

In larger systems the occurrence of a certain number of non-deterministic state transitions is the rule rather than the exception. This is basically, because the number of possible state transitions grows exponentially with system size, such that non-deterministic transitions become increasingly likely. Simulations nonetheless indicate that their relative number is always small in strongly interacting systems: Of the 2^N entries per Matrix column a fraction at most linear in N are non-vanishing. For Markov chains slightly different from those used in the simulations this can actually be proved rigorously, cf. Section 4.2. Accordingly, the internal state transitions of strongly interacting Markov chains with constrained periphery are always at least *almost* deterministic. Even though some source states may have more than one internal target state, most project to just one internal target deterministically. Therefore, the optimized Markov chains can be termed “weakly non-deterministic”: Asymptotically in system size the relative fraction of non-deterministic transitions goes to zero.

In the context of automata theory this has the following consequences: Strictly speaking, only completely deterministic Markov chains correspond with deterministic finite state automata as outlined in section 3.4. However, automata theory also defines *non-deterministic finite state automata* (NFAs), which differ from DFAs basically by the fact that given the same state and input symbol several successor states are possible (as a second less important difference they also can have more than one initial state, see Hopcroft & Ullman (1979)). This is, what we also observe in strongly interacting constrained Markov chains. Therefore, optimized constrained Markov chains correspond in general with NFAs. However, remember that the relative fraction of non-deterministic transitions is only small. Thus, the resulting NFAs are also only “weakly non-deterministic”; they are “almost” DFAs.

For clarity, we should emphasize the following: It is important to note that in the context of automata the term “non-determinism” only refers to the fact that target states must not be unique for a given state and input symbol. In the context of Markov chains the term refers to something completely different, namely the presence of truly probabilistic state transitions, such transitions that are neither 0 nor 1 but in between. However, because Markov matrices are normalized in their first argument to sum to 1, the presence of a transition with probability in $]0, 1[$ implies that at least a second transition with that property must exist for the same pair of internal and peripheral states. Thus, probabilistic non-determinism in Markov chains implies non-determinism, i.e. non-unique target states, in the induced NFAs.

N	$ \partial $	2^N	d	2^d
2	1	4	4	16
3	1	8	16	65536
3	2	8	8	256
4	1	16	48	2^{48}
4	2	16	32	2^{32}
4	3	16	16	65536

Table 1: Number of possible DFAs, 2^d , for parallel Markov chains with $|\partial|$ peripheral units and N units in total.

3.6 Ambiguity and common structure of optimized solutions

In section 4.1 we prove the plausible fact that the stochastic interaction of a system of product form $K = K'K^\partial$, cf. (7), can be written as a sum of the interaction of the peripheral Markov transition (p^∂, K^∂) and an interaction term for the internal transition (p, K') . Because the interaction of the fixed Markov transition on the periphery is constant, maximizing $I(p, K)$ is therefore equivalent to maximizing the internal interaction.

Parallel kernels K' for binary state variables now have $d := 2^N(N - |\partial|)$ independent parameters, $K^{(\nu)}(z_\nu | \omega)$, $z_\nu \in \{0, 1\}$, $\omega \in \{0, 1\}^N$, because there are $N - |\partial|$ internal units and 2^N possible source states. A parallel kernel is deterministic, if all the $K^{(\nu)}(z_\nu | z, a)$ are either 0 or 1, such that unit ν either fires or remains silent with probability 1 given any source state (z, a) . This implies, that for given N and $|\partial|$ the number of deterministic parallel Markov chains is 2^d , a number that grows unimaginably fast. For instance, for just 3 units one of which peripheral we already have 65536 possible DFAs, see Table 1.

Not all these deterministic Markov chains are local maximizers of $I(p, K)$ but it is reasonable to assume that also the number of local maximizers grows very fast with N (the more, since additional non-deterministic maximizers exist). As a consequence, the optimized chains appearing in simulations for a given N and fixed peripheral Markov transition (p^∂, K^∂) are non-unique. They are in fact highly variable. Figure 5, for instance, displays a few (out of 2^{32} in principle possible) optimized deterministic chains derived under the same conditions as the one in Fig. 2, that is $N = 4$, $|\partial| = 2$, and $K^\partial(a' | a) = 0.25$ for all $a, a' \in \Omega_\partial$. Interaction in these systems is $I(p, K) = 1.323508$, 1.35529 , and 1.331308 for A, B, and C, respectively. Apparently, all these kernels are deterministic, but the corresponding automata (not shown) are certainly not equivalent. Thus, the development of the internal state transitions is to some degree independent of the periphery.

Observe further, that all rows in the kernels contain roughly the same number of positive entries. Degenerate Markov chains, where one or more of the internal states have no incoming connections from other states appear seldom in simulations, although in principle they could maximize $I(p, K)$. Furthermore, optimized Markov chains seldom reveal state pairs z, z' such that z projects to z' independent of the current input a . This situation corresponds with completely filled blocks in kernels as displayed in Fig. 5 and 6. As a rule of thumb, the transitions in an optimized kernel are more or less randomly scattered over the whole Markov matrix.

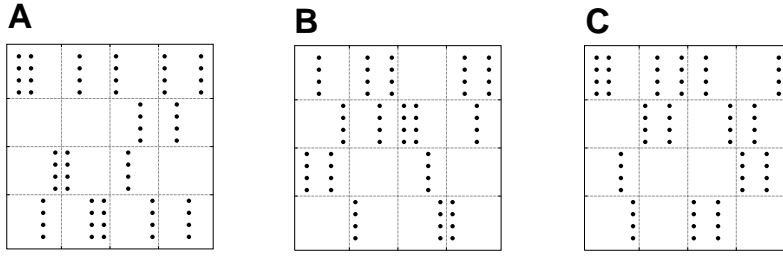


Figure 5: Different Markov kernels, $K(\omega' | \omega)$, resulting for $N = 4$, $|\partial| = 2$ and peripheral transitions all of equal probability as in Fig. 2.

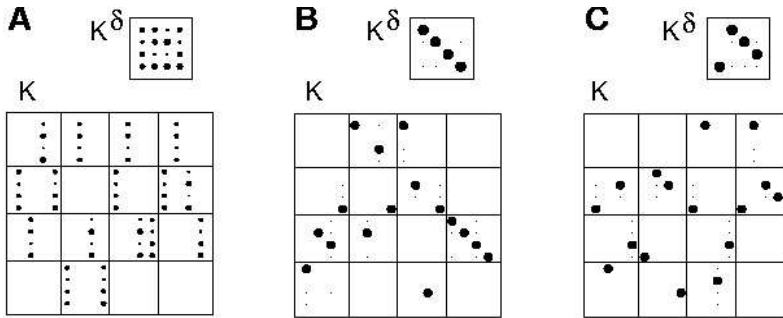


Figure 6: Optimized Markov kernels for different types of periphery as indicated by the peripheral kernels K^∂ in each frame. A) Random kernel; B) identity; C) deterministic cyclic state sequence on ∂ .

Figure 6 displays comparable simulations for various choices of K^∂ : In A K^∂ is a random kernel as in Fig. 4; in B and C the kernels are ‘deterministic’, although in order to yield a unique stationary probability distribution p^∂ , the off diagonal elements of K^∂ in B (and similar in C) were set to small positive values. As before, individual simulation runs for each of these kernels converged to a variety of different systems (not shown), with realized transitions scattered throughout the whole Markov matrices as apparent from Fig. 6. Note also, how the full kernels K reflect the Markov chain on the periphery: If an internal transition $K'(z' | z, a)$ is 1, then the whole a -th column of $K^\partial(a' | a)$ is copied into block z' of the (z, a) -th column of K . This is a consequence of the product form of $K = K' \cdot K^\partial$, cf. (7). Beside this periphery-induced fine-structure, on the level of internal transitions the coarse structure of the matrices in Figs. 5 and 6 are quite similar, again indicating a relative independence of the development of the internal state transitions from the periphery.

The described common structural features of strongly interacting Markov chains as shown in Figs. 5 and 6 will be explained in section 4 on the base of mathematical properties of the stochastic interaction measure $I(p, K)$.

3.7 Impact of periphery on final FSAs

In the previous section we have seen that the dominating qualitative features of internal state transitions in strongly interacting Markov chains are to some degree independent of the precise process on the periphery. The question remains, whether or not the peripheral Markov chain has any influence on the resulting finite state automata at all, beside determining the fine-structure for each realized internal transition. The answer is, of course, yes. However, these influences are quite subtle.

Figure 7 displays interaction values for all 16 possible deterministic Markov kernels K' for $N = 2$, $|\partial| = 1$, and different kernels K^∂ . Kernels used for that figure did not result from an optimization, but were explicitly constructed. Plus signs denote results for a peripheral kernel with all entries equal, that is, the peripheral unit is a Bernoulli process with $\text{prob}[z = 1] = 0.5$. Crosses indicate results for an ‘identity kernel’ where transitions $0 \rightarrow 0$ and $1 \rightarrow 1$ are almost 1, but small off-diagonal probabilities of 0.001 ensure that the stationary probability distribution $p^\partial = (0.5, 0.5)$ is unique. Trajectories of the peripheral unit here consist of long sequences of either zeros or ones; the small off-diagonal entries occasionally switch between both states. Squares in Fig. 7 denote a deterministic cyclic process on the periphery, i.e., trajectories $\dots 01010101 \dots$. Finally, circles in Fig. 7 represent interaction values for a parallel peripheral process with $K^{(1)}(a' = 1 | a) = (0.05, 0.2)$ which is almost a Bernoulli process with rate ≈ 0.05 but an increased probability that the output stays 1 if it was so in the previous state. Note that because we only have one peripheral unit the interaction of the periphery is always zero for the above choices, but the conditional entropies of the peripheries are $\ln 2$ (i.e., as large as possible) for the Bernoulli process with rate 0.5, 0.216273 for the specified parallel process, and zero for the identity and cyclic kernels.

At the bottom in Fig. 7 the indexes of the DFAs used in the upper part of the figure are represented binary. The respective internal Markov kernels K'_i for DFA number i are chosen such that: $K'_i(z' = 1 | \omega) = \text{bin}(i)[\omega]$ where $\text{bin}(i)[\omega]$ is the ω -th position in the binary representation of i , cf. Fig. 7. Figure 8 displays some of the DFAs as state transition graphs.

First observe in Fig. 7 that quite a few of the possible DFAs have zero interaction for all tested peripheries, i.e., DFAs 0, 3, 4, 8, and 12 to 15. As Fig. 8 reveals for DFAs $0 \equiv 0000$ and $3 \equiv 0011$ these automata are degenerate: For arbitrary input sequences, and thus, for all ergodic peripheral Markov chains, they run into trivial attractors, where the next state of each unit is always perfectly predictable from its current state alone, but independent of the current input. The internal units then generate no entropy, and the internal interaction and conditional entropy are zero.

Note further, that the spectrum of interaction values in Fig. 7 is different for different peripheral Markov chains. This indicates that the periphery indeed has an influence on which automata occur, although much of their structural properties are independent of K^∂ as argued in the previous section. The Bernoulli process, for instance, leads to a maximal interaction of $\ln 2$ for DFAs 5,6,9, and 10, shown in the bottom row in Fig. 8. In fact, in these automata, we are maximally uncertain about the next state, because each state projects to a different one for different input, and the inputs appear with equal probabilities. This is slightly different for DFAs 1, 2, 7, and 11. There, some of the

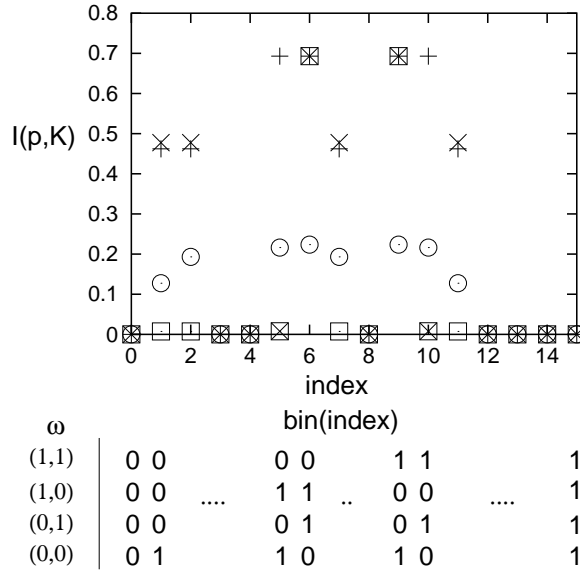


Figure 7: $I(p, K)$ for $N = 2$, $|\partial| = 1$ and all 16 possible parallel deterministic Markov kernels K' . Periphery K^∂ has been clamped to different choices (cf. also Fig. 6): plus-signs: all entries equal; crosses: identity; squares: cyclic sequence 01010...; circles: a special parallel process (see text). At the bottom binary representations of the deterministic Markov kernels K' are plotted, i.e., $K'(z' = 1 | \omega) = \text{bin}(\text{index})[\omega]$.

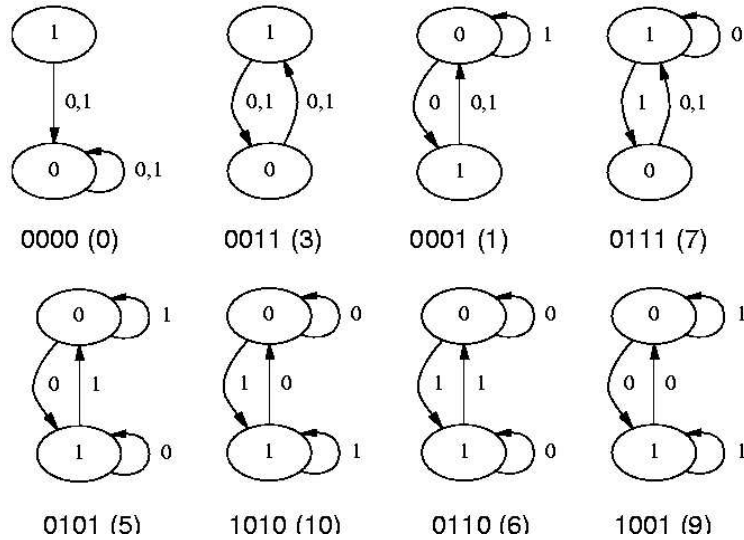


Figure 8: Some of the DFAs for $N = 2$, $|\partial| = 1$. Numbers below each transition graph are the binary (and decimal) indexes as indicated in Fig. 7.

internal state transitions are independent of the current input, but not all. Therefore, the interaction of these systems is larger than zero, but smaller than the maximally possible value $\ln 2$.

Finally, observe that the interaction of some Markov chains is high for some periph-

eries, but low or vanishing for others. DFAs 5 and 10 for instance are optimal for the maximum entropy Bernoulli process on the periphery, but for the identity and cyclic peripheral kernels their interaction is zero. As can be read from the state transition graph for DFA 5 in Fig. 8 for an identity kernel on the periphery the internal unit goes into state 0 whenever the input is 1, and vice versa. Because the peripheral process stays in each state for long times (defined by the very small off-diagonal probabilities in K^∂) the next state can therefore be almost perfectly predicted, such that the interaction vanishes. Similarly, for the cyclic periphery the internal state transitions are always $\dots 010101 \dots$, such that again the internal unit activity is perfectly predictable from its present state alone corresponding with a vanishing interaction. Conversely, for automata 6 and 9, the cyclic input $\dots 010101 \dots$ leads to internal state transitions $\dots 00110011 \dots$. Here the state transitions $z \rightarrow z'$ are all equally likely, such that we cannot predict anything about the next internal single unit state with just a one-step Markov chain (we could, however, with a two- or more step chain).

In conclusion, as shown in section 3.6 the general form of $I(p, K)$ favours optimized Markov chains, where minimizing the conditional Entropy $H(p, K)$ induces determinism, and maximizing the marginal single unit entropies leads to an “unfolding” of the chains, such that in every internal state many successor states are possible. Shown in the present section is that of the large number revealing these general structural features, the precise periphery selects chains that still lead to unpredictable firing of internal single units for the special input sequences generated by the peripheral dynamics.

3.8 Intrinsic modes of activity

Unconstrained Markov chains represent autonomous dynamical systems. As discussed in section 3.2 their dynamics is characterized by sparse graphs consisting of almost deterministic nested loops augmented by transients. Because no input is provided the corresponding dynamic attractor structures can be seen as intrinsic modes of activity of strongly interacting Markov chains in isolation of the system (cf., also Ay & Wennekers (2003)).

In contrast, the internal dynamics of a constrained Markov chain represents a non-autonomous system driven by the peripheral input. Nonetheless, for constant input, e.g., no input at all, $a = 00 \dots$, the internal state transitions of constrained optimized chains reveal the same properties as those of unconstrained chains. This is demonstrated in Fig. 9 for the same example system as displayed already in Figs. 2 and 3.

Figures 9A to D show subgraphs that draw only edges with the same label of the automaton in Fig. 3. Plot A, for instance, corresponds with the subgraph drawn bold in in Fig. 3. Apparently, activity in the decomposed graphs flows along transient trajectories into more or less complex attractor structures. The latter are all deterministic in the displayed example, but in especially in larger systems they may also contain branching points and nested loops. Furthermore, note that the approached attractor structures need not be simply connected, but may consist of several disjoint components comparable to Fig. 9D, where both, state 01 and 10 are fixed points of the system for constant input 11. Clearly, in larger systems as our simple example, the asymptotically approached attractor structures can become almost arbitrarily complex up to the point that they must stay

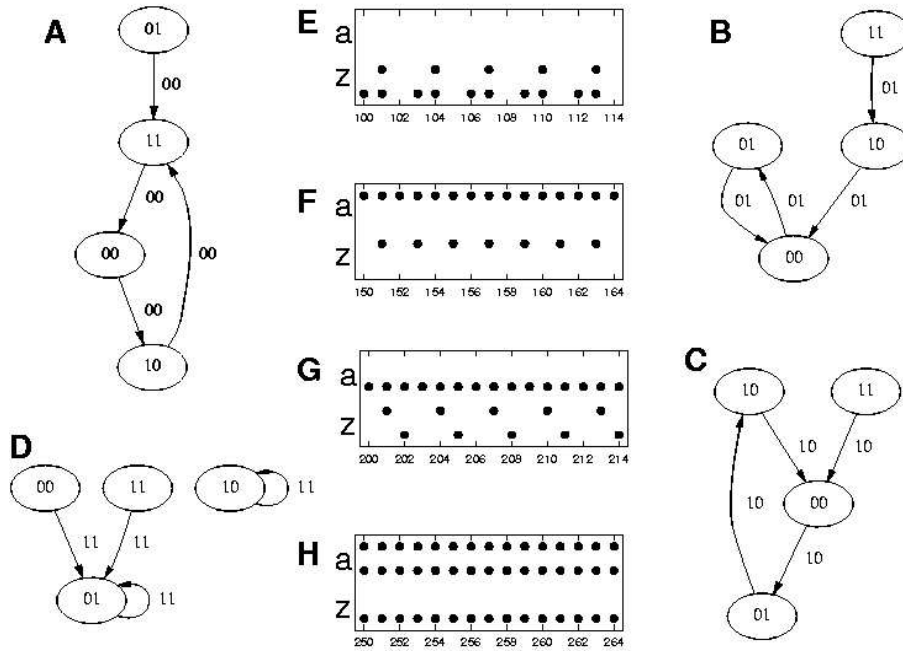


Figure 9: Same example as in Figs. 2 and 3. Plots A to D decompose the graph in Fig. 3 into different intrinsic modes for constant input 00, 01, 10, and 11. Plot A, for instance, corresponds with the subgraph drawn bold in in Fig. 3. E to H display resulting activation patterns as raster plots over time (where, as usual, units are counted from bottom to top, cf. Fig. 2).

almost deterministic.

As a consequence, activity in a strongly interacting system develops towards specific spatio-temporal firing patterns for constant input. Such characteristic “intrinsic modes of activity” are displayed in Figs. 9E to H. In this small example they consist of short repetitive firing sequences, but in larger systems comprising nested loops of various length and multiple different attractors they quickly become very complex such that firing patterns appear virtually random to a naive observer’s eye already for 5 or 6 units (cf., the examples in Ay & Wennekers (2003)). Nevertheless, they are, of course, almost deterministic and contain strong, input-specific spatio-temporal correlations.

The membrane potential dynamics of real neurons lives on a time-scale of, let’s say, roughly 10ms. Changes in the environment on the other hand are most often slower. Therefore, it seems natural to regard input into the system as quasi-stationary with respect to the internal temporal scale. This corresponds with peripheral Markov chains that switch states less often than the internal chains do. For instance, the peripheral Markov chains with identity kernel and small off-diagonal transition probabilities in some of the previous examples, reveal this property. The slowly changing input then can be seen as selecting or addressing particular subgraphs of the full state transition graph, such that the respective modes of activity become relevant for the further flow of activity. Multiple attractors, in addition, provide memory effects since specific inputs can switch between specific modes, but it is the history of the system, that determines which attractors are

selected. In that sense intrinsic modes appear as computing devices.

4 Analytical Results

In the present section we consider the optimization of strongly interacting systems mathematically. At first, section 4.1 explains the most prominent structural features of constrained optimized Markov chains on the base of analytical properties of the conditional entropy, $H(p, K)$, and the stochastic interaction measure, $I(p, K)$. Section 4.2 then proves upper bounds for the number of non-vanishing transitions and the maximum entropy of optimized Markov chains.

4.1 Additivity of H and I

We start with some instructive calculations assuming a Markov kernel of product form, cf. (7):

$$K(\omega' | \omega) = K(z', a' | z, a) = K'(z' | z, a)K^\partial(a' | a) , \quad (8)$$

where $\omega = (z, a)$, $\omega' = (z', a') \in \Omega_V$, $a, a' \in \Omega_\partial$, and $z, z' \in \Omega_{V \setminus \partial}$ as usual. Then the conditional entropy of (p, K) can be written as

$$H(p, K) = - \sum_{\omega, \omega'} p(\omega) K(\omega' | \omega) \ln K(\omega' | \omega) \quad (9)$$

$$= - \sum_{z', \underline{a}, z, a} p(z, a) K'(z' | z, a) K^\partial(\underline{a} | a) \ln K'(z' | z, a) \quad (10)$$

$$- \sum_{\underline{z}', a', z, a} p(z, a) K'(\underline{z}' | z, a) K^\partial(a' | a) \ln K^\partial(a' | a) \quad (11)$$

$$= - \sum_{z', z, a} p(z, a) K'(z' | z, a) \ln K'(z' | z, a) \quad (12)$$

$$- \sum_{a', a} \left(\underbrace{\sum_z p(z, a)}_{=p^\partial(a)} \right) K^\partial(a' | a) \ln K^\partial(a' | a) \quad (13)$$

$$= H(p, K') + H(p^\partial, K^\partial) . \quad (14)$$

In (12) and (13) we have used the normalization of the Markov kernels K^∂ and K' in their first argument. In (13), $\sum_z p(z, a) = p^\partial(a)$, where $p^\partial = K^\partial p^\partial$ is the stationary probability distribution on the periphery, because we optimize with respect to induced stationary probability distributions and the periphery is independent of the internal units.

Equation (14) shows that the total kernel entropy can be written as a sum of the conditional entropy of the periphery and that of the transition (p, K') . Since further, the entropy of the periphery is a fixed quantity, optimization of the interaction $I(p, K)$ can

only influence $H(p, K')$ in (14). With (14) the interaction measure reads

$$I(p, K) = \sum_{\nu \in V} H_\nu(p_\nu, K_\nu) - H(p, K) \quad (15)$$

$$\begin{aligned} &= \sum_{\nu \in V \setminus \partial} H_\nu(p_\nu, K_\nu) - H(p, K') + \sum_{\nu \in \partial} H_\nu(p_\nu, K_\nu) - H(p^\partial, K^\partial) \\ &= I(p, K') + I(p^\partial, K^\partial) . \end{aligned} \quad (16)$$

Equation (16) reveals that also the interaction $I(p, K)$ can be written as a sum of the interaction of the periphery and that of the Markov transition (p, K') . Again, $I(p^\partial, K^\partial)$ is constant during optimization, such that the maximization of $I(p, K)$ is actually equivalent to the maximization of $I(p, K') = \sum_{\nu \in V \setminus \partial} H_\nu(p_\nu, K_\nu) - H(p, K')$. Consider $H(p, K')$:

$$H(p, K') = - \sum_{z', z, a} p(z, a) K'(z' | z, a) \ln K'(z' | z, a) \quad (17)$$

$$= \sum_{z, a} p(z, a) \underbrace{\left(- \sum_{z'} K'(z' | z, a) \ln K'(z' | z, a) \right)}_{=H(K'(\cdot | z, a))} . \quad (18)$$

The under-braced term in (18) is obviously the Shannon-entropy generated by the internal state transitions induced by K' restricted to the fixed source state (z, a) . $H(p, K')$ then is the weighted average over these entropies, where the weights are the probabilities that the system is indeed in state (z, a) before the transition. To maximize $I(p, K')$ we should make $H(p, K')$ small. Clearly, if $K'(\cdot | z, a)$ is deterministic, that is, if there is only a single target state z' with $K'(z' | z, a) = 1$ for the given source state (z, a) , the Shannon-entropy $H(K'(\cdot | z, a))$ is zero. Thus, if all $K'(\cdot | z, a)$, $z \in \Omega_{V \setminus \partial}$, $a \in \Omega_\partial$ are deterministic, the total entropy $H(p, K')$ obtains its absolute minimum of 0. The Markov transition (p, K') then generates no entropy and, in fact, $K'(z' | z, a)$ can be interpreted as the transition table of a deterministic finite state automaton: For every internal source state z and input a there is exactly one target state z' .

Note, however, that we used parallel Markov kernels for the internal states in the simulations

$$K'(z' | z, a) = \prod_{\nu \in V \setminus \partial} K^{(\nu)}(z'_\nu | z, a) . \quad (19)$$

For K' to be deterministic, it suffices that all $K^{(\nu)}(z'_\nu | z, a)$ are either 0 or 1. Then also all $K'(z' | z, a)$ are either 0 or 1, and because a product of binary variables, i.e., the $K'(z' | z, a)$ for fixed (z, a) , is equivalent to logical AND-ing these variables, for each pair (z, a) there exists exactly one z' with $K'(z' | z, a) = 1$. Whence, $H(p, K')$ is zero for parallel kernels if they only consist of transitions with probability either 0 or 1. The product form of K in (8) then implies for the full kernels, that the a -th column inside block z, z' is just a copy of the a -th column of K^∂ , as observed and termed “periphery-induced fine-structure” in section 3.6, cf. Figs. 5 and 6.

Finally, observe that maximizing $I(p, K)$ does not only require $H(p, K')$ to be small, but in addition that the marginal entropies $H_\nu(p_\nu, K_\nu)$, $\nu \in V \setminus \partial$ are large. The impact of

this requirement is informally the following: Consider the identity on the internal states: $K'(z'|z, a) = \delta_{z',z}$. For that choice the block structure of the internal state transitions is diagonal. The kernel is also deterministic and, thus, gives a minimal zero entropy, $H(p, K') = 0$. If we would minimize $H(p, K') = 0$ alone, the identity kernel would be a global minimum. But note that it also implies $K_\nu(z'_\nu|z_\nu) = \delta_{z'_\nu, z_\nu}$ such that the internal units have constant activity and accordingly a marginal entropy of 0. They do not contribute to any information processing and add nothing to the total entropy of the system in excess of the entropy already put in on the periphery. Therefore, although the internal state transitions are deterministic and $H(p, K')$ is zero, the interaction measure is not larger than the interaction of the clamped periphery. This clearly is a type of uninteresting dynamics, which is prohibited by simultaneously minimizing $H(p, K)$ and maximizing the marginal entropies. Then, the internal state transitions “unfold” towards graph structures with many possible successor states that generate entropy by themselves. As the simulations show, these structures can still be deterministic or at least almost deterministic.

The above arguments explain heuristically the special form of the dynamics observed in the simulations: First, the convergence towards almost deterministic systems, because these maximize $H(p, K')$. Second, the fact that realized internal transitions are apparently randomly scattered throughout the optimized Markov kernels. This increases the number of possible pathways in the state transitions graphs, and therefore the unpredictability of activity of individual units, $H_\nu(p_\nu, K_\nu)$. These two factors determine the gross structure of the optimized Markov chains as discussed in section 3.6. The precise peripheral Markov chain has the further influence of weighting some of the possible pathways in the state transition graphs stronger than others, according to the probabilities with which certain input sequences appear. This makes some of the possible chains more likely, others less, but still leaves many locally optimal solutions with the described basic structural features.

4.2 Theorems

In the present section we state theorems showing that in fact *all* strongly interacting systems are weakly non-deterministic. As defined earlier a “strongly interacting system” is a local maximizer of I . On the other hand, we consider a Markov transition with constrained periphery as “weakly non-deterministic”, if for every global state ω with $p(\omega) > 0$ the number of possible internal target states is bounded by $\eta(V/\partial) + 1$ where for a subset $A \subset V$ we define

$$\eta(A) := \sum_{v \in A} (|\Omega_v| - 1).$$

For binary neurons, i.e. $|\Omega_v| = 2$ for all $v \in V$, one has $\eta(A) = |A|$, such that “weak determinism” imposes a bound linear in the number of internal units on the number of possible internal target states of the Markov kernel given any fixed global state with $p(\omega) > 0$. For strictly deterministic systems this number is exactly 1; general Markov chains, to the other extreme, can have exponentially many target states. In fact, for the unconstrained optimization of temporal interaction, $\partial = \emptyset$, we already proved the following theorem (Ay & Wennekers, 2003).

THEOREM 4.1. *Consider a probability distribution $p \in \bar{\mathcal{P}}(\Omega_V)$ and a transition kernel $K \in \bar{\mathcal{K}}(\Omega_V)$. If (p, K) is a local maximizer of I^V , then for all $\omega \in \text{supp } p$ the following bound on the support of $K(\cdot | \omega)$ holds:*

$$|\text{supp } K(\cdot | \omega)| \leq 1 + \eta(V). \quad (20)$$

For binary units the estimate (20) implies the linear bound $|\text{supp } K(\cdot | \omega)| \leq 1 + |V|$. Instead of the unconstrained optimization, we now consider a driven system. Let ∂ be a subset of the set V of neurons, the *periphery* of the system. We assume that the process on the periphery is given by the environment of the system which is fixed. We model this extrinsic process by a probability distribution $p^\partial \in \bar{\mathcal{P}}(\Omega_\partial)$ and a transition kernel $K^\partial \in \bar{\mathcal{K}}(\Omega_\partial)$. As already considered in (6), the intrinsic information processing is modeled by a transition kernel K' from Ω_V to $\Omega_{V \setminus \partial}$. Thus, we investigate the optimization of I^V restricted to the set of transition kernels K from Ω_V to Ω_V that have the product structure

$$K(z', a' | z, a) = K'(z' | z, a) K^\partial(a' | a). \quad (21)$$

This constrained optimization leads to the following generalization of Theorem 4.1.

THEOREM 4.2. *Let (p, K) be a local maximizer of the restriction of I^V to the set of transition kernels with product structure (21) and a fixed peripheral transition kernel K^∂ , and let $\omega = (z, a)$ be an element of $\text{supp } p$. Then for the intrinsic kernel K' of K we have*

$$|\text{supp } K'(z' | z, a)| \leq 1 + \eta(V \setminus \partial). \quad (22)$$

Note that we recover the estimate (20) if we set $\partial := \emptyset$ in (22). Then, formally, K^\emptyset maps the empty state ϵ onto the empty state, such that $K' = K$. Theorem (4.2) implies the following corollary on the entropy generated by a strongly interacting system.

COROLLARY 4.3. *In the situation of Theorem 4.2, the conditional entropy of the next internal state given the current global state satisfies*

$$H_{(p,K)}(X'_{V \setminus \partial} | X) \leq \ln(1 + \eta(V \setminus \partial)).$$

The proofs of Theorem 4.2 and Corollary 4.3 are given in the appendix. Informally, they imply that *all* strongly interacting systems of the form (21) – i.e., where information flows only into the system from some periphery – must be weakly non-deterministic: Given any internal state and input the number of internal target states is small (linear in system size) as compared to the possible number of states (exponential). Accordingly, the internally generated entropy grows at most logarithmically in system size ($|V \setminus \partial|$).

5 Summary and Discussion

To summarize, we have defined a measure of stochastic interaction including spatial and temporal properties of stochastic processes as the divergence of a Markov chain from its product of marginal chains. We have shown numerically and analytically that the optimization of stochastic interaction in Markov chains with clamped periphery leads to deterministic or at most weakly non-deterministic finite state automata. In these systems the dynamics prescribed on a set of input units drives the internal dynamics through (almost) deterministic state transitions. Nonetheless, the internal single unit activities in strongly interacting systems are largely unpredictable. These features are explained by the property of stochastic interaction to combine two goals: On one hand it minimizes the conditional entropy for global state transitions, but simultaneously it maximizes the single unit entropies. As a consequence, the resulting internal Markov chains are confined from arbitrary Markov kernels towards deterministic kernels, but they unfold from degenerate chains, such that in every internal state as many different target states as possible can be approached in dependence of the present input activity. This way, the recurrent internal dynamics of strongly interacting systems reveals complex internal structure, in contrast to pure feedforward networks.

From a dynamical systems viewpoint strongly interacting systems can be seen as driven or non-autonomous systems with rich internal dynamics. If the input is held constant for some time, activity flows into attractors specific for the particular input, though not necessarily unique, cf. Fig. 9. Accordingly, peripheral activity constant over a certain time can select intrinsic modes of activity, and peripheral state transitions can further switch between such internal dynamic modes. This provides the simplest way of “neural computations”, because information about the history of the system can be represented and processed. Interestingly, some experimental evidence indeed suggests the existence of brief intrinsic modes or states in cortical neural activity: As Abeles et al. (1995) have demonstrated, cortical activity in prefrontal areas of monkeys flips among quasi-stationary states of several ten to hundred milliseconds duration defined by short-time firing rate patterns of simultaneously recorded neurons. Similar phenomena appear in our network for slowly varying input patterns, if the intrinsic dynamics is forced into different modes over time. Gat et al. (1997) have further shown that the mentioned state flips can be well segmented by Hidden Markov models suggesting that intrinsic modes can be switched on a fast time-scale although they persist themselves on a longer scale. It would be interesting to determine the stochastic interaction comprised by these experimentally observed Markovian systems and compare it with complete randomness or order under various behavioural conditions.

Aertsen et al. (1989) defined “functional connectivity” of a set of neurons with reference to short-time correlations in their mutual firing patterns. In experimental data, these correlations were shown to change rapidly over time, with the interpretation that neurons dynamically form varying subgroups of interacting cells, also phrased “functional cell assemblies”. Interestingly, in our optimized systems correlations would change similarly if the network activity is driven through different intrinsic modes, but they are constant, determined by the transiently approached attractor, in each particular mode. This way,

our approach may provide an explanatory base for the complex correlation dynamics found in experiments.

A further aspect of cortical neural activity seems important at that point: This is the presence of repetiting firing patterns with interspike intervals up to the order of tens to hundreds of milliseconds. Abeles et al. (1993a) have shown that such long-lasting “synfire-patterns” appear reliable and behavior-dependent in multiple electrode recordings of monkeys performing simple behavioral tasks. In light of the largely stochastic firing of single units this observation is highly surprising (Abeles, 1990). The classical “synfire chain model” explains these long-time correlations by volleys of synchronized activity that propagate repeatedly along the same, i.e., deterministic, neural pathways (Abeles, 1990; Abeles et al., 1993b). Activity in our optimized Markov chains reveals quite similar properties: Single unit activity is virtually random, but the whole state transitions are largely deterministic and proceed along nested repetitive loops. So, a network dynamic can be globally deterministic even if every single neuron’s activity looks virtually random. In fact, on the background of neural assemblies and associative memories already in Wennekers (1998) we have demonstrated that the classical synfire chain model can be extended in a simple and straightforward way to implement arbitrary deterministic and non-deterministic finite state automata. Wennekers & Ay (2003) furthermore argue that synfire chain type activation patterns appear naturally under the assumption that the brain maximizes temporal interaction. Attractor models of brain function on the other hand reveal only small stochastic interaction.

We further mention a relation of our work to a series of papers by Tononi, Sporns, and Edelman (Sporns et al., 2000; Tononi et al., 1994, 1999). They considered the segregation and integration of neurons into functional ensembles based on several different measures for complexity: Shannon-entropy, spatial interaction also termed “integration” in Tononi et al.’s work, and two further measures that account for information flow between partitions of a set of units. The “integration” measure is equivalent to our stochastic interaction restricted to stationary probability distributions. Tononi et al. compared structural features of systems that optimize one or the other complexity measure. As a main result they found that in particular their partition-based measures lead to networks with distinct structural characteristics such as clustered connectivity and a short wiring length (cf. also Murre & Sturdy (1995) for an interesting complementary approach). Whence, the neurons organize into mutually segregated subgroups, with strong internal interactions. The spatial “integration” measure, however, usually leads to systems where most cells are bound into a single strongly interacting cluster. A difference between our work and Tononi et al.’s is, of course, that stochastic interaction as used in the present work, is based on spatial *and* temporal interactions. Our example systems, therefore, show a rich internal and input dependent dynamics and are better described in space-time – i.e., as the intrinsic modes – rather than in space alone. We have to leave mathematical conceptualizations of these issues to future work.

Our principle of temporal information maximization complements Linkser’s Infomax principle for stationary input-output relations in layered feedforward systems. Linkser’s work pointed out surprising links between two previously unrelated and even distant areas of scientific research: Information maximization and the structure of visual receptive

fields. The principle of Temporal Infomax as developed in the present work presents a reasonable extension of Linsker's classical Infomax to the spatio-temporal domain. And again, also here it turns out that Information Maximization can suddenly be linked to a previously completely unrelated area of research: The theory of computing machines. The possibility of grounding the development of automata-like computational structures in neural systems on information theoretic optimization principles seems as appealing as Linsker's observation that such principles may guide the organization of sensory hierarchies. Both principles, of course, need further experimental evaluation.

Acknowledgement

Main parts of this work have been performed when both authors were at the Max Planck Institute for Mathematics in the Sciences. Nihat Ay thanks the Santa Fe Institute for hosting him during the final work at the present paper.

A Proofs

Now we come to the proof of Theorem 4.2. It is based on the following lemma, which we have proved in Ay & Wennekers (2003):

LEMMA A.1. *Let $\bar{\Delta}$ be a d -dimensional closed simplex in a real vector space and ext its set of extreme points. For each subset $ext' \subset ext$, $\Delta(ext')$ denotes the open face of $\bar{\Delta}$ with the extreme points ext' , and we have the stratification*

$$\bar{\Delta} = \bigsqcup_{\emptyset \neq ext' \subset ext} \Delta(ext').$$

For a point $x \in \bar{\Delta}$, $\text{supp } x$ is the subset of ext defined by $x \in \Delta(\text{supp } x)$. Now consider an affine subspace V of $\text{aff } \bar{\Delta}$ that is given by r linear equations:

$$V = \{x \in \text{aff } \bar{\Delta} : x \text{ satisfies the } r \text{ given linear equations}\}.$$

If a point $x_0 \in \mathcal{C} := V \cap \bar{\Delta}$ locally maximizes a strictly convex function $f : \mathcal{C} \rightarrow \mathbb{R}$, then

$$|\text{supp } x_0| \leq d + 1 - \dim V \leq \min\{r, d - \dim \mathcal{C}\} + 1. \quad (23)$$

PROOF OF THEOREM 4.2. We fix the local maximizer (p, K) of the interaction I^V , an $\omega \in \text{supp } p$, and define the simplex

$$\begin{aligned} \bar{\Delta} &:= \bar{\Delta}(p, K^\partial, K', \omega) \\ &:= \left\{ (p, K^\partial, L') \in \bar{\mathcal{P}}(\Omega_V) \times \bar{\mathcal{K}}(\Omega_\partial) \times \bar{\mathcal{K}}(\Omega_{V \setminus \partial} | \Omega_V) : \right. \\ &\quad \left. L'(\cdot | \sigma) = K'(\cdot | \sigma) \text{ for all } \sigma \in \Omega_V, \sigma \neq \omega \right\} \\ &\subset \mathbb{R}^{\Omega_V} \times \mathbb{R}^{\Omega_\partial \times \Omega_\partial} \times \mathbb{R}^{\Omega_V \times \Omega_{V \setminus \partial}}. \end{aligned}$$

This set can naturally be identified with $\bar{\mathcal{P}}(\Omega_{V \setminus \partial})$ by the map $\bar{\Delta} \rightarrow \bar{\mathcal{P}}(\Omega_{V \setminus \partial})$, $(p, K^\partial, L') \mapsto L'(\cdot | \omega)$. Now we define the convex subset

$$\mathcal{C} := \{(p, K^\partial, L') \in \bar{\Delta} : L'_v = K'_v \text{ for all } v \in V \setminus \partial\},$$

which can be represented as the intersection of $\bar{\Delta}$ with an affine subspace of $\mathbb{R}^{\Omega_V} \times \mathbb{R}^{\Omega_\partial \times \Omega_\partial} \times \mathbb{R}^{\Omega_V \times \Omega_{V \setminus \partial}}$ that is given by $\eta(V \setminus \partial)$ equations. In order to apply Lemma A.1, we have to prove that the interaction I^V is strictly convex on \mathcal{C} . This part of the proof follows exactly the lines in Ay & Wennekers (2003) for the unconstrained case and is therefore not repeated here. Lemma A.1 then implies

$$|\text{supp } K'(\cdot | \omega)| \leq 1 + \eta(V \setminus \partial).$$

□

PROOF OF COROLLARY 4.3. This follows directly from Theorem 4.2. □

References

- Abbott, L. & Sejnowski, T.J., eds. (1999) *Neural Codes and Distributed Representations*. MIT press, Cambridge MA.
- Abeles, M. (1991) *Corticonics: Neural circuits of the cerebral cortex*. Cambridge: Cambridge University Press.
- Abeles, M., Bergman, H., Gat, I., Meilijson, I., Seidemann, E., Tishby, N., & Vaadia, E. (1995) Cortical Activity Flips Among Quasi Stationary States. *Proc.Natl.Acad.Sci. (USA)* 92, 8616–8620.
- Abeles, M., Bergman, H., Margalit, E., & Vaadia, E. (1993) Spatio-temporal firing patterns in frontal cortex of behaving monkeys. *J.of Neurophysiol.* 70, 1629–1643.
- Abeles, M., Vaadia, E., Bergman, H., Prut, Y., Headman, I., & Slovin, H. (1993) Dynamics of Neuronal Interactions in the Frontal Cortex of Behaving Monkeys. *Concepts in Neuroscience* 4, 131–158.
- Aertsen, A. (ed.) (1993) *Brain Theory*. Elsevier Science Publishers, Amsterdam.
- Aertsen, A. M. H. J., Gerstein, G. L., Habib, M. K., & Palm, G. (1989) Dynamics of Neuronal Firing Correlation: Modulation of “Effective Connectivity”. *J. Neurophysiol.* 61, 900–917.
- Ay, N. (2001) *Information Geometry on Complexity and Stochastic Interaction*. MIS-Preprint Series 95/2001. Submitted.
- Ay, N. (2002) Locality of Global Stochastic Interaction in Directed Acyclic Networks. *Neural Computation* 14 (12), 2959–2980.

- Ay, N., & Wennekers, T. (2003) Dynamical properties of strongly interacting Markov chains. *Neural Networks* 16, 1483–1497.
- Barlow, H. (2001) Redundancy reduction revisited. *Network: Computation in Neural Systems* 12, 241–253.
- Bell, A. J., & Sejnowski, T. J. (1995) An information maximization approach to blind separation and blind deconvolution. *Neural Computation* 7, 1129–1159.
- Cover, T.M., & Thomas, J.A. (1991) *Elements of Information Theory*. Wiley Series in Telecommunications. New York: Wiley-Interscience.
- Cutler, A., & Breiman, L. (1994) Archetypical analysis. *Technometrics* 36, 338–347.
- Dayan, P., & Abbott, L.F. (2001) *Theoretical Neuroscience*. MIT-Press, Cambridge, MA.
- Eckhorn, R. (1999) Neural mechanisms of scene segmentation: Recordings from the visual cortex suggest basic circuits for linking field models. *IEEE Transactions on Neural Networks*, 10, 464–479.
- Fairhall, A. L., Lewen, G. D., Bialek, W., & van Steveninck, R. R. D. (2001) Efficiency and ambiguity in an adaptive neural code. *Nature* 412, 787–792.
- Gat, I., Tishby, N., & Abeles, M. (1997) Hidden Markov modelling of simultaneously recorded cells in the associative cortex of behaving monkeys. *Network - Computation in Neural Systems* 8, 297–322.
- Gerstein, G. L., Bedenbaugh, P., Aertsen, A. M. H. J. (1989) Neuronal assemblies. *IEEE Transactions on Biomedical Engineering* 36, 4–14.
- Hopcroft, J.E., & Ullman, J. D. (1979) *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley, Reading, MA, 1979.
- Hertz, J., Krogh, A., & Palmer, R. G. (1991) *Introduction to the theory of neural computation*. Addison Wesley.
- Lee, T. W., Girolami, M., Bell, A. J., et al. (2000) A unifying information-theoretic framework for independent component analysis. *Comput.Math.Appl.* 39, 1–21.
- Li, Z. & Arick, J. J., (1994) Toward a theory of the striate cortex. *Neural Computation* 6, 127–146.
- Linsker, R. (1986a) From Basic Network Principles to Neural Architecture: Emergence of spatial-opponent cells. *Proceedings of the National Academy of Sciences (USA)* 83, 7508–7512.
- Linsker, R. (1986b) From Basic Network Principles to Neural Architecture: Emergence of orientation-selective cells. *Proceedings of the National Academy of Sciences (USA)* 83, 8390–8394.

- Linsker, R. (1986c) From Basic Network Principles to Neural Architecture: Emergence of orientation columns. *Proceedings of the National Academy of Sciences (USA)* 83, 8779–8783.
- 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.
- Martignon, L., Deco, G., Laskey, K., Diamond, M., Freiwald, W., & Vaadia, E. (2000) Neural Coding: Higher-order temporal patterns in the neurostatistics of cell assemblies. *Neural Computation*, 12, 2621–2653.
- Murre, J. M. J., & Sturdy, D. P. F. (1995) The connectivity of the brain: Multi-level quantitative analysis. *Biological Cybernetics* 73, 529–545.
- Nakahara, H., & Amari, S. (2002) Information geometric measure for neural spike trains. *Neural Comput.* 14, 2269–2316.
- Palm, G., & Aertsen, A., (editors) (1986) *Brain Theory*. Springer, Berlin.
- Penev, P. S. & Atick, J. J. (1996) Local feature analysis: A general statistical theory for object representation. *Network: Computation in Neural Systems* 7,477–500.
- Rieke, F., Warland, D., Ruyter van Steveninck, R., & Bialek W. (1998) *Spikes: Exploring the Neural Code*. Cambridge: MIT Press.
- Singer, W., & Gray, C.M. (1995) Visual feature integration and the temporal correlation hypotheses. *Annual Review of Neuroscience*, 18, 555–586.
- Sporns, O., Tononi, G., & Edelman, G. M. (2000) Connectivity and complexity: the relationship between neuroanatomy and brain dynamics. *Neural Networks* 13, 909–922.
- Studený, M., & Vejnarova, J. (1998) The multiinformation function as a tool for measuring stochastic dependence. In Jordan M.I. (ed.) 1998. *Learning in Graphical Models*, Dordrecht: Kluwer, 1998.
- Tononi, G., Sporns, O., & Edelman, G. M. (1999) Measures of redundancy and degeneracy in biological networks. *Proc.Natl.Acad.Sci. (USA)* 96, 3257–3262.
- Tononi, G., Sporns, O., & Edelman, G. M. (1994) A measure for brain complexity: Relating functional segregation and integration in the nervous system. *Proc.Natl.Acad.Sci. (USA)* 91, 5033–5037.
- Wennekers, T. (1998) *Synfire Graphs: From Spike Patterns to Automata of Spiking Neurons*. Technical Report # 98-08, Faculty for Computer Science, University of Ulm.
- Wennekers, T., & Ay, N. (2003) Spatial and temporal stochastic interaction in neuronal assemblies. *Theory Biosci.* 122, 5–18.
- Wennekers, T., Sommer, F., & Aertsen, A. (editors) (2003) *Neural Assemblies*. Special Issue of *Theory in Biosciences* 112, 1–104.